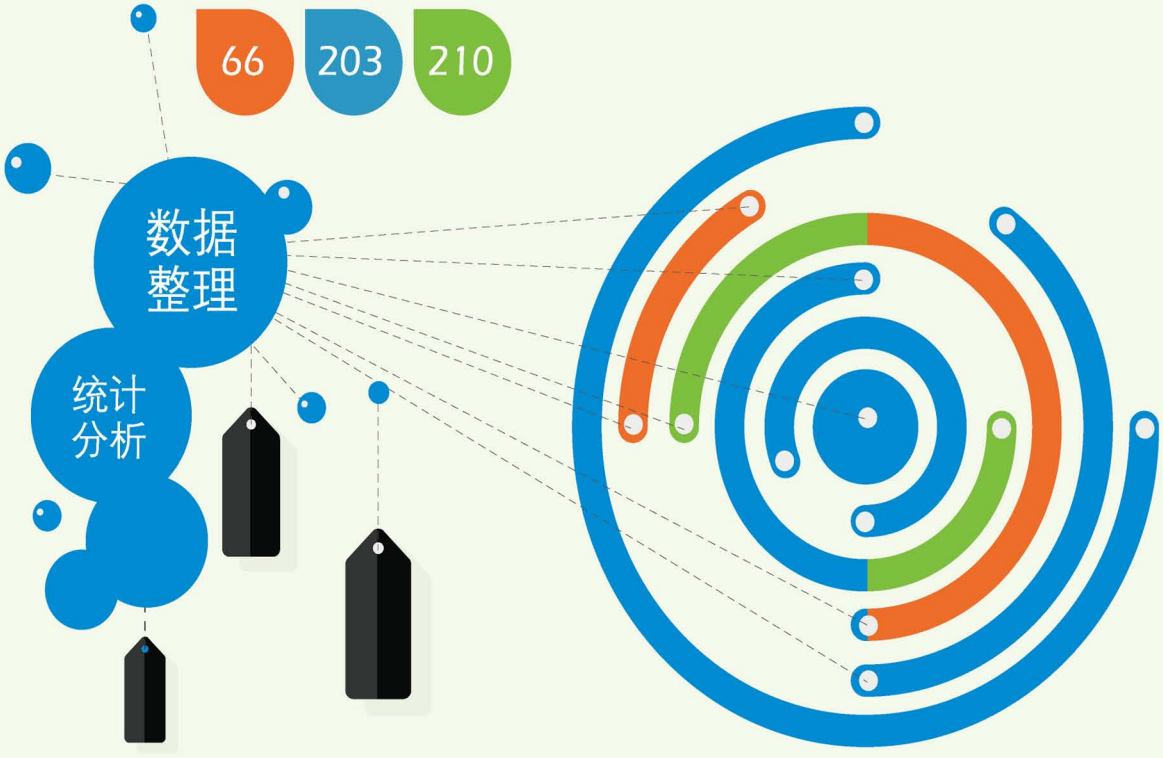
Broadview
www.broadview.com.cn

Excel与数据分析（第3版）

宇传华 编著



十年沉淀，为你呈现

结合前两版的读者反馈，保留精华，补充变化，充分展示用Excel做统计的优势！

電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn



/ 作者简介 /

宇传华，武汉大学公共卫生学院教授，博导，副院长。美国华盛顿大学（西雅图）生物统计学访问学者，中山大学医学统计学博士后，第四军医大学卫生统计学博士，同济医科大学硕士、学士。教育部高等学校（医学人文素质）教学指导委员会委员，湖北省卫生统计与信息学会副会长、中国卫生信息学会理事、《中国卫生统计》与《公共卫生与预防医学》等杂志编委。发表(译)著作29部，论文80余篇。主持有国家自然科学基金等项目。主要从事全球疾病负担定量方法与应用、诊断试验评价、数据挖掘与软件开发等研究。

Excel 与数据分析

（第 3 版）

宇传华　编著

電子工業出版社
Publishing House of Electronics Industry
北京・BEIJING

内 容 简 介

本书由浅入深地介绍了 Excel 数据整理、统计分析、高级应用及电脑实验的方法。利用 Excel 函数，编制了近 100 个统计分析模板，如样本含量与检验功效的估计、Meta 分析、诊断实验评价、灰色模型预测、圆形分布分析等，这些模板由 Excel“数据分析”工具是无法实现的。此外，本书还利用 Excel 函数，制作了约 50 个统计学电脑实验，这些实验可帮助读者理解抽象的统计学概念，有利于统计学教学。本书配套的光盘操作简单，读者只需复制或键入数据，随即便可获得统计分析或电脑实验结果。

本书适用于和数据打交道的各界人士，可作为科研人员、管理人员、医务工作者、大中专院校师生的参考书或教材，尤其适用于学过或打算学习统计学的有关人员。

未经许可，不得以任何方式复制或抄袭本书之部分或全部内容。
版权所有，侵权必究。

图书在版编目（CIP）数据

Excel 与数据分析 / 宇传华编著. —3 版. —北京：电子工业出版社，2013.7
ISBN 978-7-121-20739-6

Ⅰ. ①E… Ⅱ. ①宇… Ⅲ. ①表处理软件 Ⅳ.①TP391.13

中国版本图书馆 CIP 数据核字（2013）第 133525 号

责任编辑：付 睿
印 刷：北京丰源印刷厂
装 订：三河市鹏成印业有限公司
出版发行：电子工业出版社
北京市海淀区万寿路 173 信箱 邮编 100036
开 本：787×1092 1/16 印张：36 字数：806 千字
印 次：2013 年 7 月第 1 次印刷
印 数：4000 册 定价：79.00 元（含光盘 1 张）

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

服务热线：（010）88258888。

著者说明

本书主要章节由宇传华主笔完成，以下人员在本书（包括前面的两个版本）及 Excel 程序的编写等方面做出了较大贡献。

人员名单

武汉大学	王震坤　杨子娟　李　鹏　连　肖 姚　聪　孙　敏　季　洁　李　妍 夏　欣　魏娜娜　郝玉伟
华中科技大学	王家春
北京法马苏提克医药有限公司	方龙
首都医科大学	祝慧萍
广西中医药大学	李京
腾讯公司	林嵩艺
深圳市卫生人口和计生委员会	董国营
广州军区武汉总医院	文芸
深圳市宝安区妇幼保健院	钟文明
上海市普陀区疾病预防控制中心	蒋丽丽
上海市嘉定区疾病预防控制中心	马飞飞
广东省检验检疫局	颜杰
中山大学	张晋昕

前　言

本书第 1 版是 2002 年出版的《Excel 与数据分析》，第 2 版是 2009 年出版的《Excel 统计分析与电脑实验》，前面版本中的很多内容，特别是统计分析程序，至今仍非常简单实用，第 3 版继承了上述版本的所有优点。这 10 年来，每当读者以各种方式对本书发出赞许之声，或者以各种方式告之本书已售罄的消息时，我就感觉到自己责任的重大，有一种强烈的努力编写好本书的欲望。

第 3 版是以 Excel 2010 为基础编写的。从 Excel 2003 过渡到 Excel 2010 的开始几天，操作起来有些不习惯，甚至有点觉得它过于花哨与烦琐。但经过一段时间的适应后，发觉 Excel 2010 改进真的很大，从“功能选项卡”到“命令组”，再到每一个“命令按钮”，层次更加清晰，选项内容也更加丰富；将“文件”选项卡独立于 Excel 工作表之外，更方便了对整个文件的操作。Excel 更稳定了，Excel 函数增多了，兼容性更强了，函数计算精确度更高了，Excel 图形可独立保护与撤销保护等新功能都提示用户从使用 Excel 2003 过渡到使用 Excel 2010 的正确性与迫切性。

第 3 版所配光盘中的每一个统计学分析模板与电脑实验程序均在 Excel 2010 中运行正常。从数量上看，自编的 Excel 程序（或界面）第 1 版为 66 个，第 2 版是 203 个，现在的第 3 版是 210 个。绝大多数程序只需键入数据，结果便自动在同一工作表内产生，具有即改即可见的效果。

选用 Excel 做统计分析的理由

与标准统计学软件 SAS、SPSS 等相比较，Excel 的优势在于：①强大的数据自动填充功能，这使得在 Excel 工作表中输入数据变得相当简单；②方便的数据汇总与数据透视分析功能，可快速获得整个工作表，甚至整个工作簿数据的探索性统计分析结果；③灵活的单元格绝对引用与相对引用功能，使得大量类似的 Excel 公式输入变得简单，很多情况下只要键入一个公式，通过拖放方式便可完成该单元格周围其他单元格公式的生成；④完美的图表内置格式，使得统计图形既制作方便，又美观好看；⑤Excel 的“数据分析”工具可完成大多数常用的统计分析，对于较复杂的统计学方法可简单利用 Excel 内置的函数自编程序完成；⑥正规渠道购买的电脑内大多数预装有微软 Office 办公软件，且为中文界面，使用它进行统计学分析无版权问题。

本书特色

本书从实例出发，易学易懂。所配光盘内包括有大量实例数据与程序，很多情况下，只需在编好的 Excel 程序中键入有关数据或参数，随即便可获得计算结果。

所配光盘程序均为开放代码，读者通过书中有关提示，结合光盘文件可以自编新的程序，解决更复杂的统计学分析难题。部分 Excel 程序采用了工作表保护，目的是防止 Excel 工作表内重要数据或计算公式发生不经意的改动。如果需要去除保护，只需选择“审核→撤销工作表保护”命令，输入密码“0”即可。

本书光盘程序是在 Windows 7 操作系统下，以 Excel 2010 为基础，屏幕分辨率为 1366×768 像素下制作完成的。

本书内容

本书共分为数据整理、统计分析、高级应用和电脑实验 4 篇。在数据整理篇中，较为全面系统地介绍了数据的编辑、查询、建立子数据清单、数据透视表与透视图分析等方法。在统计分析篇中，利用 Excel 图表向导功能制作了多种常见统计图形；利用 Excel 内置函数编制了大量统计分析模板，如卡方检验、正态分布等优度拟合检验、Fisher 精确概率法、非参数检验、Ridit 分析、简单相关与回归分析、多重相关与回归分析、加权 logistic 回归分析等。在高级应用篇中编制了样本含量与检验功效的估计、寿命表、诊断试验评价的样本含量估计、ROC 曲线分析、Meta 分析、TOPSIS 法、层次分析法等统计模块。在电脑实验篇中，利用“开发工具→插入→ActiveX 控件→滚动条”制作滚动条，结合 Excel 内置函数和内置图表，制作了大量电脑实验，如正态分布、二项分布、Poisson 分布等十余个概率分布电脑实验；z 分布、t 分布、F 分布、卡方分布等抽样分布电脑实验、中心极限定理、随机现象、假设检验、置信区间等共计近 50 个电脑实验程序。

第 3 版还新增加了灰色预测模型（第 12.6 节）、圆形分布分析（第 12.8 节）、Bland-Altman 法（第 16.6 节）等统计学方法与 Excel 实现程序。

对于“实用型”用户，可直接利用有关程序进行统计学分析；对于“高级”用户，可根据书中提示，结合光盘的相应文件，对相应工作表略加修改，完成更复杂的计算。

本书编者

本书第 1、2、3 版的所有编写人员已在“著者说明”页列出。

在第 3 版的修改编写过程中，第 1～5 章以及附录 A、B、C、D、E 由王震坤负责，第

6～10 章由杨子娟负责，第 11～16 章由李鹏负责，第 17～20 章由连肖和姚聪负责。此外，王震坤在编写第 12.6 节，以及联络其他编者等方面做出了较大贡献。

限于作者水平有限，时间仓促，书中错谬之处在所难免，恳请各位同行、专家和广大读者批评指正。

宇传华（E-mail：yuchua@163.com）
武汉大学
2013 年 6 月

目　　录

数据整理篇

第 1 章　Excel 基本知识 ····· 2

1.1　Excel 的安装 ····· 3
1.2　Excel 工作界面简介 ····· 3
1.3　鼠标与键盘的操作 ····· 9
1.4　单元格区域与整个工作表的选取 ····· 12
1.5　工作表重命名与复制副本 ····· 13
1.6　工作表的格式化 ····· 14
1.7　工作表的打印 ····· 19

第 2 章　公式与函数 ····· 21

2.1　公式的建立 ····· 22
2.2　函数的结构 ····· 23
2.3　公式中的运算符 ····· 24
2.4　运算符的优先顺序 ····· 26
2.5　单元格的混合引用 ····· 26
2.5.1　相对引用 ····· 27
2.5.2　绝对引用 ····· 27
2.5.3　混合引用 ····· 27
2.5.4　A1 引用样式 ····· 28
2.5.5　应用实例 ····· 28
2.6　编辑公式常用的快捷键 ····· 30
2.7　编辑公式时常见的错误信息 ····· 30
2.8　公式保护与撤销保护的技巧 ····· 33
2.9　常用数学函数应用举例 ····· 34
2.9.1　常用数学函数 ····· 35
2.9.2　数组运算 ····· 39
2.9.3　矩阵的运算与求解线性方程组 ····· 40

2.10 创建自定义函数……43
2.11 获取函数帮助信息……44

第 3 章 数据清单的建立与利用……45

3.1 数字的单元格格式……46
3.2 数据的有效性设置……50
3.3 建立数据清单……54
3.4 数据“记录单”的作用……56
3.5 数据的自动填充……57
3.5.1 简单数据的自动填充……57
3.5.2 等比数列数据的填充……59
3.5.3 公式的自动填充……60
3.5.4 智能标记的应用……60
3.6 数据的分列……61
3.7 添加与编辑批注……63
3.8 特殊数学符号的键入……64
3.9 数据的导入与导出……66

第 4 章 数据的编辑与查询……68

4.1 窗口的冻结与拆分……69
4.2 数据的剪贴与删除……70
4.3 数据的查找与替换……71
4.4 行列转置与选择性粘贴的其他妙用……72
4.5 将连续型变量划分等级……75
4.6 年龄的计算与当前日期的生成……76
4.7 数据的跨工作表操作……76
4.8 多个工作表的合并计算……78
4.9 查询与新建子数据清单……81
4.9.1 排序……81
4.9.2 自动筛选……82
4.9.3 高级筛选……85
4.9.4 选择不重复的记录……86
4.9.5 人工筛选……86
4.10 两个类似文件的比较……88

第 5 章　分类汇总与透视分析 …… 90
5.1　分类汇总 …… 91
5.2　数据透视表 …… 94
5.2.1　创建数据透视表 …… 95
5.2.2　数据透视表的结构及其工具栏 …… 96
5.2.3　由数据透视表获得均数标准差 …… 98
5.2.4　产生列联表 …… 100
5.2.5　创建子数据清单 …… 106
5.2.6　设置报告格式与更新数据 …… 106
5.2.7　在数据透视图表中创建公式 …… 107
5.3　数据透视图 …… 108
5.3.1　创建数据透视图 …… 109
5.3.2　调整数据透视图 …… 109
5.3.3　将数据透视图更改为静态图 …… 111
5.4　数据透视的其他应用 …… 113
5.4.1　多个工作表的数据透视 …… 113
5.4.2　数据透视表网页 …… 113

统计分析篇

第 6 章　统计图表 …… 116
6.1　统计表的制作 …… 117
6.2　统计图的结构及特点 …… 118
6.3　Excel 标准图表类型简介 …… 119
6.4　创建统计图的一般步骤 …… 121
6.5　统计图的编辑与修饰 …… 124
6.6　绘制统计图举例 …… 127
6.6.1　分段条图 …… 127
6.6.2　误差条图 …… 129
6.6.3　百分条图 …… 131
6.6.4　饼图 …… 132
6.6.5　复合条饼图 …… 133
6.6.6　线图与半对数线图 …… 135

6.6.7 雷达图 136
6.6.8 气泡图 138
6.6.9 箱图 140
6.6.10 人口金字塔图 143
6.6.11 三角函数制作线图 146
6.6.12 曲面图 147

第 7 章 统计描述 149

7.1 数据分析工具的调用 150
7.2 统计学中的数据类型 151
7.2.1 数据类型的辨析 152
7.2.2 三类数据的计算 153
7.2.3 三类数据的转化 153
7.3 区间数据的统计学描述 154
7.3.1 区间数据的频数表与直方图 154
7.3.2 集中位置指标 160
7.3.3 离散程度指标 163
7.3.4 数据分析工具：描述统计 169
7.4 名义数据的统计描述 170
7.4.1 COUNTIF 函数清点计数 171
7.4.2 率的标准化法 171
7.4.3 动态数列 173

第 8 章 区间数据的假设检验 176

8.1 假设检验的一般步骤 177
8.2 样本均数与总体均数的比较 178
8.2.1 已知样本均数和标准差 178
8.2.2 已知样本原始数据 179
8.3 配对资料的比较 180
8.4 两个样本均数的比较 181
8.4.1 两样本的总体方差齐性检验 181
8.4.2 总体方差齐同的 *t* 检验 183
8.4.3 总体方差不齐的 *t* 检验 185
8.4.4 大样本的 *z* 检验 186

8.5 多样本方差 Bartlett 齐性检验 ····· 187
8.6 单因素方差分析 ····· 188
8.6.1 单因素方差分析工具 ····· 188
8.6.2 单因素方差分析图示 ····· 190
8.6.3 无原始数据的单因素方差分析 ····· 191
8.6.4 单因素方差分析的多重比较 ····· 191
8.7 双因素方差分析 ····· 193
8.7.1 无重复双因素方差分析 ····· 193
8.7.2 有重复双因素方差分析 ····· 195

第 9 章 名义与有序数据的假设检验 ····· 198

9.1 卡方检验 ····· 199
9.1.1 四格表资料的χ^2检验 ····· 199
9.1.2 四格表配对资料的χ^2检验 ····· 203
9.1.3 四格表资料的 Fisher 精确概率检验 ····· 205
9.1.4 行×列表资料的χ^2检验 ····· 208
9.2 频数分布的拟合优度检验 ····· 210
9.2.1 正态分布 ····· 210
9.2.2 二项分布 ····· 213
9.2.3 Poisson 分布 ····· 215
9.2.4 均匀分布 ····· 216
9.3 非参数检验 ····· 217
9.3.1 配对资料 Wilcoxon 符号秩检验 ····· 218
9.3.2 两独立样本非参数检验 ····· 221
9.3.3 多样本 Kruskal-Wallis 秩和检验 ····· 225
9.3.4 多个相关样本的 Friedman 检验 ····· 228
9.4 Ridit 分析 ····· 230

第 10 章 相关与回归 ····· 235

10.1 简单相关与回归 ····· 236
10.1.1 Excel 函数实现简单相关与回归 ····· 238
10.1.2 数据分析工具实现简单相关与回归 ····· 243
10.1.3 添加趋势线实现直线回归分析 ····· 247
10.2 曲线回归 ····· 248

10.2.1 对数曲线拟合 ····· 249
10.2.2 乘幂曲线拟合 ····· 251
10.2.3 指数曲线拟合 ····· 252
10.2.4 多项式曲线拟合 ····· 253
10.3 非线性回归 ····· 255
10.3.1 对数曲线拟合 ····· 255
10.3.2 乘幂曲线拟合 ····· 256
10.3.3 指数曲线拟合 ····· 257
10.3.4 多项式曲线拟合 ····· 257
10.4 多重相关与回归 ····· 258
10.4.1 多重相关与多重协方差 ····· 258
10.4.2 多重回归分析 ····· 260
10.4.3 用函数实现多重回归与相关 ····· 261
10.5 加权 logistic 回归 ····· 263

高级应用篇

第 11 章 变量求解与方案优选 ····· 268

11.1 单变量求解 ····· 269
11.1.1 目标搜索 ····· 269
11.1.2 求解非线性方程 ····· 270
11.2 规划求解 ····· 271
11.3 方案管理器 ····· 274
11.4 常见的方案优选方法 ····· 277
11.4.1 确定型决策 ····· 277
11.4.2 非确定型决策 ····· 278
11.4.3 风险型决策 ····· 281
11.5 综合评价方法 ····· 282
11.5.1 TOPSIS 法 ····· 282
11.5.2 层次分析法 ····· 285

第 12 章 预测分析 ····· 288

12.1 自相关与自回归 ····· 289
12.1.1 线性时序趋势 ····· 289

12.1.2 DW 统计量……290
12.1.3 自相关……291
12.1.4 自回归……292
12.2 移动平均法……294
12.3 指数平滑法……295
12.3.1 数据分析工具——指数平滑……296
12.3.2 指数平滑的初始值……297
12.3.3 求解最佳平滑常数……298
12.4 季节预测法……299
12.4.1 哑变量多重回归……300
12.4.2 AR（4）模型……303
12.4.3 季节指数法……304
12.5 Markov 预测法……306
12.6 灰色预测模型……310
12.6.1 GM(1,1)模型的预测与评价方法……310
12.6.2 灰色模型预测实例……312
12.7 傅利叶分析……314
12.8 圆形分布分析……316
12.8.1 角度变换……316
12.8.2 平均角及其假设检验……317
12.8.3 圆形分布分析实例……318
12.8.4 两个或多个样本平均角的比较……320

第 13 章 寿命表……324

13.1 现时寿命表……325
13.2 去死因寿命表……330
13.3 健康期望寿命表……334

第 14 章 随机化分组与样本含量估计……337

14.1 随机化分组……338
14.1.1 完全随机设计……338
14.1.2 配对设计……339
14.1.3 配伍组设计……340
14.1.4 三个随机化分组程序介绍……342

14.2 样本含量的估计 343
14.2.1 Nomogram 图示法 345
14.2.2 假设检验的样本含量估计 346
14.2.3 置信区间的样本含量估计 353
14.3 检验功效的估计 357
14.4 流行病学研究的样本含量与检验功效 362
14.4.1 病例对照研究 362
14.4.2 队列研究 365
14.5 方差分析等方法的样本含量与检验功效 366
14.5.1 方差分析的检验功效 366
14.5.2 方差分析的样本含量 367
14.5.3 其他方法的样本含量与检验功效 368

第 15 章 高级科研统计图与 Meta 分析 369

15.1 统计地图 370
15.2 P-P 图与 Q-Q 图 373
15.2.1 创建绘图数据 373
15.2.2 绘制正态 P-P 图 374
15.2.3 绘制正态 Q-Q 图 375
15.3 生存曲线 375
15.3.1 Kaplan-Meier 法计算生存率及其标准误 376
15.3.2 绘制生存曲线 377
15.4 重复测量数据的曲线图 378
15.5 Meta 分析统计量的置信区间图 379
15.5.1 Meta 分析的计算公式 380
15.5.2 连续资料的 Meta 分析 383
15.5.3 离散资料的 Meta 分析 387
15.5.4 相关系数资料的 Meta 分析 391
15.5.5 *P* 值的 Meta 分析 392

第 16 章 诊断试验评价 394

16.1 常用统计指标 395
16.2 ROC 曲线原理及其绘制方法 399
16.2.1 有序分类资料 ROC 分析 400

16.2.2 连续资料 ROC 分析……403
16.2.3 值小为阳性的 ROC 分析……404
16.3 诊断试验评价的 Meta 分析……406
16.3.1 SROC 分析……406
16.3.2 ROC 面积的 Meta 分析……412
16.4 样本含量估计……415
16.4.1 灵敏度与特异度的置信区间估计……415
16.4.2 ROC 曲线下面积的置信区间估计……416
16.4.3 固定 FPR 的灵敏度置信区间估计……417
16.4.4 单个准确度指标的假设检验……418
16.4.5 两灵敏度或特异度的比较……420
16.4.6 ROC 曲线下面积的比较……421
16.5 Kappa 值估计……422
16.6 Bland-Altman 法……425

电脑实验篇

第 17 章 概率分布……428

17.1 正态分布……429
17.1.1 采用公式制作正态分布电脑实验……429
17.1.2 正态曲线下的面积……431
17.1.3 正态分布综合电脑实验……433
17.2 二项分布……435
17.2.1 概率质量函数……435
17.2.2 累计分布函数……436
17.2.3 二项分布电脑实验……436
17.2.4 样本率及率的标准误……438
17.3 Poisson 分布……438
17.3.1 概率质量函数与累计分布函数……439
17.3.2 Poisson 分布电脑实验……439
17.4 其他连续分布……440
17.4.1 对数正态分布……441
17.4.2 均匀分布……441
17.4.3 指数分布……442

17.4.4 Weibull 分布 442
17.4.5 Beta 分布 444
17.4.6 Gamma 分布 444
17.5 其他离散分布 445
17.5.1 均匀分布 445
17.5.2 超几何分布 446
第 18 章 随机抽样与抽样误差 448
18.1 随机数发生器与随机函数 449
18.1.1 随机数发生器 449
18.1.2 随机数函数 454
18.2 有放回抽样与无归还抽样 455
18.2.1 “抽样”分析工具 455
18.2.2 无归还抽样 456
18.3 随机抽样电脑实验 456
18.3.1 从正态总体中随机抽样 457
18.3.2 从非正态总体中随机抽样 461
第 19 章 假设检验的概率分布与统计用表 466
19.1 z 分布 467
19.1.1 标准正态分布曲线下面积 467
19.1.2 标准正态分布分位数 469
19.1.3 标准正态分布的其他应用 469
19.1.4 z 分布电脑实验 472
19.2 t 分布 473
19.2.1 t 分布的图形及其分布特征 474
19.2.2 t 分布曲线下面积与检验界值 476
19.2.3 t 分布电脑实验 477
19.3 F 分布 479
19.3.1 F 分布的图形及其分布特征 480
19.3.2 F 分布曲线下面积与检验界值 481
19.3.3 F 分布电脑实验 483
19.4 χ^2 分布 484
19.4.1 χ^2 分布的图形及其分布特征 484

19.4.2 χ^2 分布曲线下面积与检验界值 …… 486
19.4.3 χ^2 分布电脑实验 …… 486

第 20 章 其他电脑实验与置信区间估计 …… 489

20.1 随机现象 …… 490
20.2 置信区间 …… 491
20.2.1 置信区间的电脑实验 …… 492
20.2.2 置信区间的估计 …… 493
20.3 假设检验的两型错误 …… 498
20.4 不同检验方法所导致结论的差异 …… 499
20.5 *t* 检验的两型错误 …… 501
20.6 *t* 与 *t'*检验 …… 502
20.7 一般 *t* 检验做方差分析两两比较导致的 I 型错误 …… 503
20.8 样本含量对假设检验的影响 …… 504
20.9 卡方检验公式校正的意义 …… 506
20.10 回归参数意义理解 …… 508
20.11 寿命表电脑实验 …… 509

附录 A Excel 函数 …… 510

附录 B 概率基本知识 …… 536

附录 C Excel 2010 宏 …… 538

附录 D Excel 2010 的数据分析新功能 …… 546

附录 E Excel 2010 VBA 应用 …… 551

参考文献 …… 554

数据整理篇

第 1 章　Excel 基本知识

第 2 章　公式与函数

第 3 章　数据清单的建立与利用

第 4 章　数据的编辑与查询

第 5 章　分类汇总与透视分析

第 1 章

Excel 基本知识

为了便于您较顺利地阅读后面章节的内容，本章将简要介绍 Excel 的基本术语、基本操作等有关知识。如果您以前使用过 Excel，则可将本章视作一个简单的复习。即使如此，您依旧可能发现一些您以前未曾注意到的技巧和方法。本书将以 Excel 2010 为主进行介绍，由于一直以来 Excel 的核心功能并未发生很大变化，故本书内容绝大多数也兼容先前的 Excel 版本。

1.1 Excel 的安装

安装 Excel，如安装 Microsoft Office 2010（即 Office 2010）软件时，建议采用“从本机运行全部程序”方式（Excel 2003 是自定义为“完全安装”方式）。因为本书后面有关章节中提到的 Excel 功能，只有以“完全安装”方式安装后才能实现，如使用：① 微软的 Symbol 字体；② Excel“数据分析”工具等。

如果以上提到的功能在运行 Excel 2010 时没有，采用后面相关章节提示的选择“文件→加载项”仍无法显现时，建议采用下列步骤修改安装方式：双击“setup.exe”可执行文件，进入安装界面后选择“添加或删除功能”，单击“继续”，然后在“安装选项”界面下单击“Microsoft Office”左侧下三角按钮，选择“从本机运行全部程序”命令。单击“继续”完成配置。

如果不希望安装微软拼音输入法，可以在“安装选项”界面下单击“Office 共享功能”前面的加号，然后单击“Microsoft Office 微软拼音输入法”左侧下三角按钮，选取“不可用”后再单击“继续”完成配置。

1.2 Excel 工作界面简介

Excel 以工作簿（book）形式构成文件，一个工作簿文件可包含多个工作表（sheet）、图表（chart）、宏指令表（macro）、VB 设计对话框（dialog），这些都可以同时存在于同一个工作簿文件内。当保存工作簿时，会把工作簿中的工作表、图表、宏指令表、VB 设计对话框一并保存在同一个工作簿文件中。

Excel 启动后，如果调入一个工作簿文件（如 book1），可获得类似于图 1-1 的工作界面，按图从上到下的顺序，Excel 工作界面包含如下几项内容：标题栏、功能区、编辑栏、工作表、工作表标签、滚动条、状态栏等。下面分别介绍它们的作用。

（1）标题栏（Title bar）

标题栏告诉用户正在运行的程序名称和正打开的文件名称。如图 1-1 所示，标题栏显示“book1.xlsx - Microsoft Excel”，表示此窗口的应用程序为 Microsoft Excel，在 Excel 中打开的当前文件名为“book1.xlsx”。标题栏右侧的  分别是 Excel 工作界面的最小化按钮、最大化和还原按钮以及关闭按钮。

（2）功能区（Ribbon）

从 Office 2007 起，Excel 的界面和以往版本相比有很大的变化，最大的变化就是使用功能区代替了以前版本中的菜单和工具栏。下面我们详细介绍 Excel 2010 的功能区。

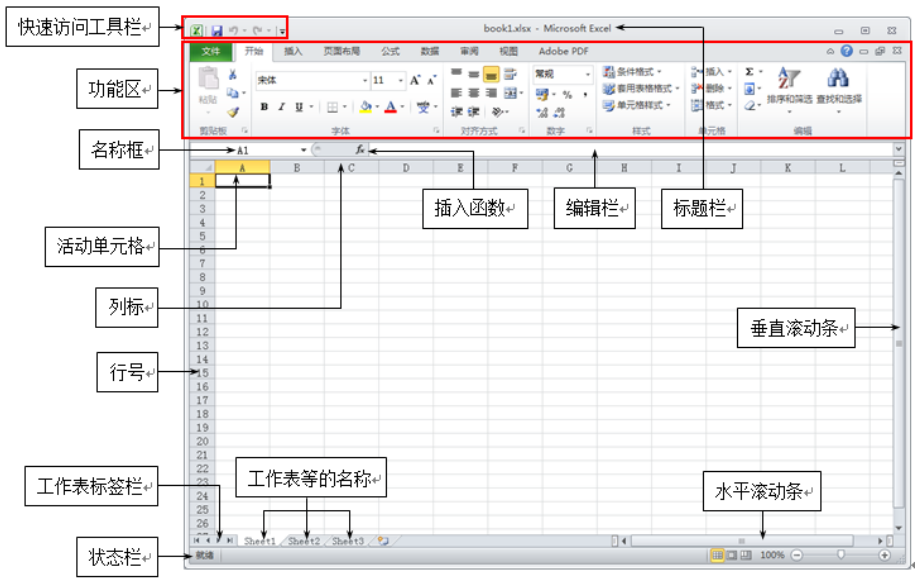


图 1-1　Excel 2010 工作界面

选项卡：位于功能区的顶部。标准的选项卡为“开始、插入、页面布局、公式、数据、审阅、视图”（如图 1-2a 所示），默认的选项卡为“开始”选项卡，用户可以在想选择的选项卡上单击来选择该选项卡。



图 1-2a　标准的选项卡

注意：①虽然“文件”按钮与各个选项卡共享了同一空间，但它实际上并不是一个选项卡。单击“文件”按钮将显示新的后台视图（见图 1-2b），可在该视图中对文档执行保存、打印等操作，此外对工作簿是否保护、是否显示某些选项（如表格的“自动套用格式”）、调用“数据分析”工具等均在此处设置与选取。②如果想要隐藏功能区以增大工作表视图，则只需双击任意选项卡或单击选项卡右侧“功能区最小化”图标即可。当需要再次使用功能区时，只需单击一个选项卡或单击选项卡右侧“展开功能区”图标，即可重新显示功能区。读者也可以按下 Ctrl+F1 组合键来切换功能区的显示和关闭。

命令组：位于每个选项卡内部。例如，“开始”选项卡中包括“剪贴板、字体、对齐方式、数字、样式、单元格、编辑”等命令组，相关的命令组合在一起来完成各种任务。

“命令”的表现形式为图形、菜单或按钮，被安排在命令组内。



图 1-2b “文件”视图

温馨提示：为了使本书操作的叙述简洁明了，本书将采用“→”替代大部分语言描述，例如采用句子②来替代句子①的表达方式如下：

① 单击功能区“开始”选项卡中“对齐方式”命令组内的“自动换行”命令。

② 单击“开始→对齐方式→自动换行”。

或

① 单击功能区“开始”选项卡中“编辑”命令组内的“清除”命令按钮旁边的下三角按钮，在弹出的菜单里选择“清除格式”命令。

② 单击“开始→编辑→清除▼→清除格式”。

鉴于以往版本中为读者所熟知的工具栏（Tool bar）已基本被整合到各选项卡的每一个命令组之中，为方便读者，对各选项卡概述如下。

“开始”选项卡：在大部分时间里，都可能需要在单击“开始”选项卡的情况下进行工作。该选项卡包含基本的剪贴板命令、格式命令、样式命令、插入和删除行或列的命令，以及各种工作表编辑命令。

“插入”选项卡：单击该选项卡可在工作表中插入需要的任何内容，如插入表格、插图、图表、符号等。

“页面布局”选项卡：该选项卡包含的命令可影响工作表的整体外观，包括主题以及一些与打印有关的设置等。

“公式”选项卡：使用此选项卡可执行插入公式、命名单元格或区域、访问公式审核工具，以及控制 Excel 执行计算的方式等操作。

“数据”选项卡：此选项卡提供了 Excel 中与数据相关的命令，包括数据的获取、链接、

编辑数据以及对其排序和筛选等。

为方便理解本书后面会讲到的操作，请读者按如下步骤在“数据”选项卡内获得“分析”命令组：选择“文件→选项→加载项”，在其右侧区域选择“分析工具库”，单击最下一行中部的“转到”按钮，在弹出的“加载宏”对话框内选择“分析工具库”以及“规划求解加载项”。这样“数据”选项卡内就会生成新的名为“分析”的命令组，其中包含有“数据分析”及“规划求解”两个命令。

“审阅”选项卡：该选项卡包含的工具用于检查拼写、翻译单词、添加注释，以及保护工作表。使用“更改”命令组，可以完成保护工作表、保护工作薄、审阅修订的功能。

“视图”选项卡：该选项卡包含的命令用于控制有关工作表显示的各个方面。例如调整显示的比例，是否全屏幕显示，是否显示“编辑栏、网格线、标题”，窗口冻结与拆分，查看宏与录制宏等。

“加载项”选项卡：如果加载了旧工作簿或者加载了会自定义菜单或工具栏的加载项，则会显示此选项卡。Excel 2010 中不再提供某些菜单和工具栏，而是在“加载项”选项卡中显示这些用户界面自定义。

“开发工具”选项卡：默认情况下不会显示这个选项卡。它包含的命令对程序员（比较高级的 Excel 用户）十分有用。为方便理解本书后面会讲到的操作，请读者按如下步骤获得该选项卡：选择“文件→选项→自定义功能区”，在其右侧区域选中“开发工具”旁边的复选框，最后单击“确定”完成配置。

（3）快速访问工具栏（Quick access toolbar）

Excel 2010 窗口中的快速访问工具栏用于放置最常使用的命令按钮，用户可以将功能区中的命令按钮和“文件”视图中的命令添加到快速访问工具栏中。以向 Excel 2010 快速访问工具栏添加“粘贴”命令为例，其操作方法如下。

打开 Excel 2010 窗口，在“开始”选项卡的“剪贴板”命令组中右键单击“粘贴”按钮，然后在打开的快捷菜单中选择“添加到快速访问工具栏”命令，如图 1-3（左侧）所示。

这样便可以在 Excel 2010 快速访问工具栏中看到“粘贴”命令按钮。如果需要将其从快速访问工具栏删除，可以在快速访问工具栏中右键单击该按钮，并在打开的快捷菜单中选择“从快速访问工具栏删除”命令即可，如图 1-3 右侧所示。

（4）编辑栏（Formula bar）

编辑栏在数据编辑整理时最常用。先选取活动单元格，在编辑栏中用户可以通过输入、复制、剪切、粘贴等方式编辑整理文字、数字或公式等。按 Enter 键，或单击×（取消）按钮，或单击 √（输入）按钮结束编辑。

编辑栏的左侧为名称框，显示活动单元格所在的位置（列标＋行号）。例如，图 1-1 中的活动单元格处于工作表的第 1 列（即列标为 A）第 1 行（即行号为 1），所以名称框显示

为 A1。当使用公式时，名称框将显示函数的名称（如 SUM 等）。名称框右侧有一个倒三角符号，在使用公式时，可供选择的函数有 SUM、AVERAGE、IF、HYPERLINK、COUNT、MAX、SIN、SUMIF、PMT、STDEV，其他函数。

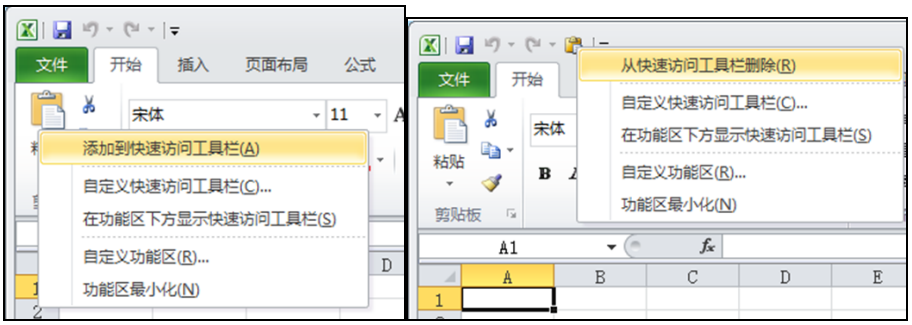


图 1-3 “添加到快速访问工具栏”（左）与“从快速访问工具栏删除”（右）

名称框和编辑栏之间一般只有一个按钮，即插入函数 fx 按钮（见图 1-1；Excel 97、Excel 2000 为=号，包括 Excel 2010 在内的 Excel 2003 以后版本修改为 fx ）。当鼠标从活动单元格移向并单击编辑栏时，中间便出现 3 个按钮图标，从左至右分别是×（取消）按钮、√（输入）按钮和 fx （插入函数）按钮。取消编辑的公式或文字等用×，确认编辑的公式或文字等用√，为了在活动单元格插入函数，用 fx 。

Excel 默认工作界面中显示有编辑栏，但如果由于某种原因使工作界面缺少编辑栏，则可在“视图”选项卡的“显示”命令组中单击“编辑栏”开关命令，选取“编辑栏”。

（5）工作表与工作表名称

工作表（sheet）是一个由列和行组成的表格。列标（column headings，如 A、B、C）和行号（row headings，如 1、2、3）分别用字母和数字表示。Excel 2003 之前的列标从左到右范围为 A～IV（即 A、B、…、Z；AA、AB、…、BA、BB、…、IV）（共 256 列），行号由上自下范围为 1～65536。每张工作表的最大容量为：65536 行×256 列＝16777216 单元格，这就是说，一张工作表最多可容纳 65536 个观察个体（记录）和 256 个变量（或字段名）（见图 1-4 左侧）。

Excel 2010 大大增加了数据容量，每张工作表的最大容量为：1048576 行×16384 列＝17179869184 单元格，这就是说一张工作表最多可容纳 100 万余个观察个体（记录）和 1.6 万余个变量（或字段名）（见图 1-4 右侧）。这样 Excel 2010 在原来版本的基础上，其行、列、单元格数量分别为 Excel 2003 及以前版本的 16、64、1024 倍。

若从其他数据库导入数据到 Excel 中，如果数据超过了 Excel 的容量，则多余数据将被 Excel 截除。所以导入其他数据到 Excel 时应特别谨慎（特别是对 Excel 2003 及以前版本）。

工作表名称也叫工作表标签，系统默认的名称有 Sheet1、Sheet2、Sheet3 等，用户可

双击鼠标左键（或单击鼠标右键，选择“重命名”命令）来修改工作表的名称。在一般情况下，Excel 同时显示 3 个工作表（其名称分别为 Sheet1、Sheet2、Sheet3），如果需要增加工作表，可在某一工作表名称处单击鼠标右键，选择“插入（I）…”或“移动或复制工作表（M）”命令，创建或复制一张工作表。一般来说，一个工作簿可显示 255 张工作表，如 Sheet1、Sheet2、…、Sheet255。

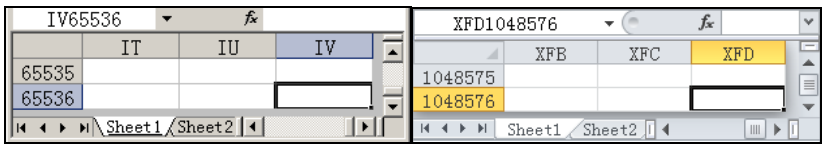


图 1-4　Excel 2003 以前版本（左）与 Excel 2010（右）的最大数据容量

当一个工作簿文件中有很多工作表时，为了查看或选取其他工作表，可利用工作表名称系列左边的一组滚动按钮 ◀ ◀ ▶ ▶ 。在 4 个滚动按钮中，按第 1、4 个按钮可分别查看最左端、最右端的工作表。按第 2、3 个按钮可分别向左、向右逐步移动工作表。工作界面显示的工作表为当前工作表，当前工作表的名称背景色为白色，其他为灰色。

（6）单元格与单元格区域

每一个列、行坐标所指定的位置称为单元格（cell），这是工作表的最小单位。单元格采用“列标 + 行号”命名，如第 B 列第 25 行单元格名称为“B25”，图 1-4（左侧）中活动单元格（active cell）名称为“IV65536”，图 1-4（右侧）中活动单元格名称为“XFD1048576”。在单元格中，用户可以键入汉字、字母、数字、符号等内容。

由若干个连续的单元格构成的矩形区域称为单元格区域（range），使用矩形左上角和右下角的两个单元格地址标识这一区域，中间用 : 相隔。例如，以 A1 为左上角、C4 为右下角的 12 个单元格组成的区域，可用 A1:C4 表示。图 1-5 中的活动单元格区域可表示为 C2:D6。

按住鼠标左键不放，然后拖动鼠标至预定单元格，可选取单元格区域；如果按住 Ctrl 键，然后用鼠标选取单元格区域，可一次选取多个单元格区域。

（7）滚动条（Scroll bar）

当工作表很大，超出一个屏幕显示时，可以使用工作表窗口右边及右下边的滚动条，使窗口在整张表上移动进行查看。

也可以通过修改常用工具栏中“显示比例框”的参数来扩大或缩小整个工作表的显示比例。

（8）状态栏（Status bar）

状态栏位于 Excel 窗口底部。它的左端是信息区，显示了 Excel 的当前工作状态，如出现“就绪”、“编辑”等信息；右端是键盘状态区，显示若干按键的开关状态，如按“Caps Lock”

键时，状态栏中便显示“大写”；按“Num Lock”键时，状态栏中便显示“数字”。

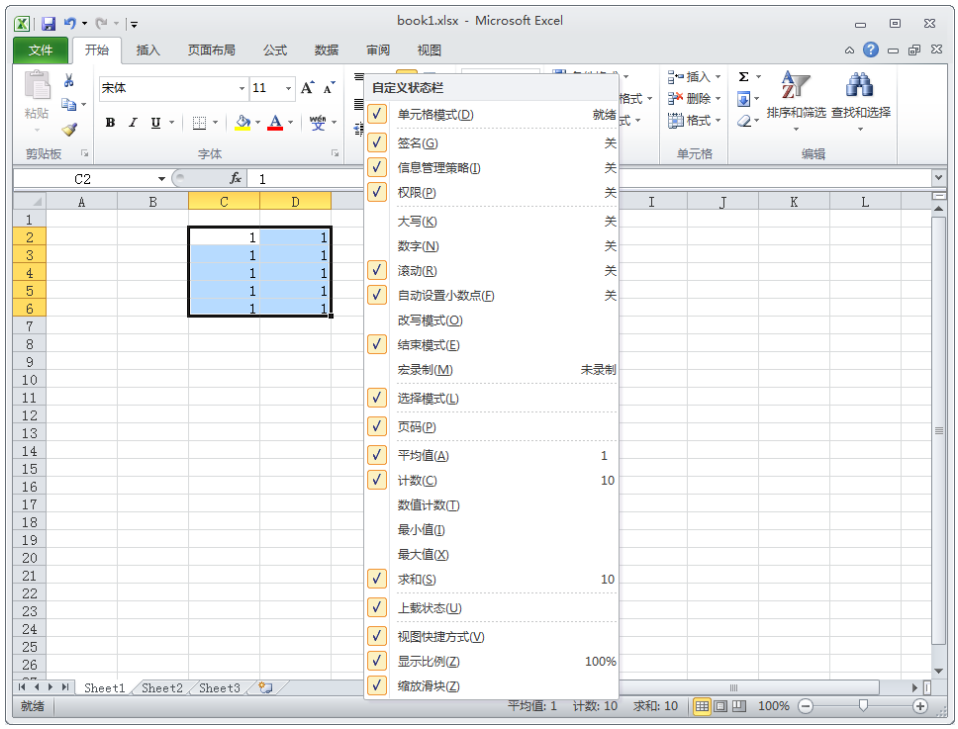


图 1-5　状态栏的计算器功能

此外，当选择一个或多个含有数据的单元格区域时，Excel 将具有计算器的功能，可以进行平均值、计数、计数值（即只对数值进行计数，不对文字、符号等计数）、最大值、最小值、求和的计算。有时需要调整显示情况，可在状态栏上任何地方单击鼠标右键，其他计算的选项菜单便会出现，如图 1-5 所示。

Excel 2010 的最右侧默认显示“显示比例”和“缩放滑块”。其计算结果除默认显示所选区域的求和结果外，也默认显示平均值、计数等计算结果。

1.3　鼠标与键盘的操作

Excel 的操作既可使用键盘，也可使用鼠标。但一般使用鼠标更快捷方便，如改变工作表的行高、列宽，移动、复制单元格的数据，自动填充数据或公式，以及图表的编辑和修饰等操作，使用鼠标可进行任意拖放，十分方便、高效。当然，也有使用键盘更为方便的情况，如输入文字、数值。

1. 常见鼠标操作

当使用鼠标时，屏幕上会出现一个指针。它相当于键盘的光标，而且随着指向对象的不同和操作的不同，鼠标指针会显示出不同的形状（见表 1-1）。

表 1-1　常见Excel鼠标形状

名　称	形　状	说　明
空心十字		Excel 工作表区域最常显示的形状
实心十字		当鼠标指针指向当前（活动）单元格或单元格区域右下角（填充柄）时出现。此时可按住鼠标左键进行拖放，完成文字、数据或公式的自动填充
插入指针		当鼠标位于编辑栏、文本框、处于编辑状态时出现。此时可输入文字、数字等
双向箭头		当鼠标位于窗口、浮动工具栏、图形对象等的边缘或角上时出现。此时可改变窗口、工具栏以及其他对象的形状或大小
十字双向箭头		当鼠标指针指向行号或列标的分界线时出现。此时可改变行高或列宽
十字箭头		当鼠标指针指向某个图形对象或固定工具栏的移动柄时变为该形状。此时可移动图形对象或固定工具栏的位置

鼠标的操作主要有：单击、双击、右击和拖放。

①单击（click）：按鼠标左键一次。常用于光标定位或选定与取消某个选项。

②双击（double-click）：快速连续按鼠标左键两次，常用于激活某个对象。例如，在打开文件对话框时，双击某个文件可直接打开该文件；在工作簿窗口中双击某个工作表名称（标签），可以重命名该工作表；双击某个单元格，则可以直接进入单元格的编辑状态。

③右击（right-click）：按鼠标右键一次。根据操作环境的不同和指向对象的不同，右击后弹出的快捷菜单也可能有所不同。例如，右击工作表中的某个单元格，可以弹出有关剪切、复制、粘贴、插入、删除、清除和设置单元格格式等命令的快捷菜单。如果右击工具栏，则可以弹出显示所有的工具栏选项的快捷菜单。由于快捷菜单中的命令都是针对当前环境和当前对象的，所以使用起来十分方便、快捷。在 Excel 中应用相当频繁。

④拖放（drag）：按住鼠标左键不放，移动光标至某一特定位置，再放开左键。拖放操作实现的功能较多。例如：

- 选定某个单元格区域，可用鼠标从该单元格区域的左上角拖放到右下角来完成。
- 特有数据或公式的自动填充。
- 移动、复制对象或改变对象的形状、大小。

2. 键盘操作

在 Excel 中，键盘操作主要用于单元格数据或公式等的输入与编辑。常见的 Excel 键盘操作见表 1-2。注意表中按键内的所有“+”不是指从键盘中键入“+”号，而是指除按住

前面的按键外，还需继续按后面的按键。

表 1-2 常见的Excel键盘操作

输入数据			单元格或编辑栏中		
按 键	功 能		按 键	功 能	
Enter	输入方向	完成单元格输入并下移一行	=	公式编辑	建立公式开始
Shift+Enter		完成单元格输入并上移一行	Alt+=		插入“自动求和”公式
Tab		完成单元格输入并右移一列	F4		输入公式时，变换单元格的相对、绝对、混合引用符号“$”
Shift+Tab		完成单元格输入并左移一列	Ctrl+ `（撇号，位于左上侧）		在显示数值与显示公式间转换（用于公式的显示）
Home		移到行首	Ctrl+ ‘（单引号，位于大键盘中偏右）		将当前单元格上方单元格中的公式复制到当前单元格或编辑栏（公式不变）
箭头键		向上、下、左、右移动一个字符	F9	计算方式	计算所有打开工作簿中的所有工作表
Esc	取消输入	取消单元格输入	Ctrl+Alt+F9		计算活动工作簿中的所有工作表
BackSpace		删除插入点左边的字符	Shift+F9		计算活动工作表
Delete		删除插入点右边的字符	Ctrl+Shift+Enter	矩阵	获得一组数据而不是单个数值（数组运算）
Ctrl+Delete		删除插入点到行末的文本	Ctrl+;	时间输入	输入当前日期
Ctrl+D	填充方式	向下填充	Ctrl+Shift+;		输入当前时间
Ctrl+R		向右填充	Ctrl+K	其他	插入超链接
Ctrl+Enter		用当前输入内容填充选定的单元格区域	Ctrl+Shift+ ‘		将当前单元格上方单元格中数值复制到当前单元格或编辑栏
Shift+F2	其他	编辑单元格批注	F2		编辑当前单元格，并将插入点放到行末
Ctrl+F3		定义名称	F7		显示“拼写检查”对话框
Alt+Enter		在同一个单元格内转行			

表 1-2 中的按键很有用，且无法单独使用鼠标键来实现，如实现数组（或矩阵）的 Ctrl+Shift+Enter 组合键，实现字母、数字在同一个单元格内转行的 Alt+Enter 组合键等。

其他键盘操作有：剪切采用 Ctrl+X 组合键，复制采用 Ctrl+C 组合键，粘贴采用 Ctrl+V 组合键；光标移到工作表使用范围的末尾（右下角）采用 Ctrl+End 组合键，光标移到工作表的 A1 单元格（左上角）采用 Ctrl+Home 组合键；选择编辑菜单按 Alt+E 组合键等。

3. 鼠标、键盘配合操作

还有些操作需要用键盘配合鼠标完成。常用的按键有：Shift 键、Ctrl 键、Alt 键。

（1）Shift 键

①绘出正方形或圆形：在绘图操作时，按住 Shift 键再用矩形或椭圆工具，可分别绘制正方形或圆形（微软 Word 的操作也是如此）。

②选择单元格区域：单击某单元格时，该单元格成为活动单元格，当单击另一个单元格时，该单元格成为新的活动单元格。如果按住 Shift 键，再单击另一个单元格，则两个单元格之间的所有单元格均成为活动单元格区域。

③选择工作表名称（标签）区域，操作与②相同。

（2）Ctrl 键

①对于不连续多个单元格区域的选取，可采用按住 Ctrl 键，再用鼠标选取若干单元格或单元格区域来实现。

②对于不相邻多个工作表的选取，操作与①相同，即按住 Ctrl 键，再用鼠标选取不同的工作表名称（标签）。

③在绘图操作时，按住 Ctrl 键再用（上、下、左、右）箭头键，可实现目标对象的微细调节（微软 Word 的操作也是如此）。

（3）Alt 键

相当于按 F10 键，作用是激活菜单栏。

Alt+Enter 组合键可实现同一单元格内的文字、数字转行，编辑含有度量衡单位的统计表时常用。

4. 功能区操作

几乎所有操作均可通过功能区完成，对于初学者特别有用，但往往影响操作速度。

以上是 Excel 中很典型的几种操作方法。在实际工作中，应根据情况尽量选择最简洁的操作方法，以提高操作速度。

1.4 单元格区域与整个工作表的选取

1. 单元格区域的选取

复制数据、输入数据、建立公式等操作经常要选取单元格区域。既可以选取一个单元格区域，也可以选取多个彼此分离的单元格区域。

选取一个单元格区域可按下列步骤进行。

① 将鼠标指针指向所要选取的单元格区域左上角。

② 按住鼠标左键，然后沿着对角线从左上角拖动鼠标到右下角。

③松开鼠标左键。

为了选取不相邻单元格区域，需要首先按下 Ctrl 键，然后重复按照上述选取一个单元格区域的方法进行多次选取。

2. 整行或整列的选取

选取整行（整列）的方法比较简单，只需将鼠标指针移动到所要选取行的行号（列的列标）处，然后单击即可。按下 Ctrl 键可进行多行或多列的选取。

在整行被选取的情况下，向上或向下拖动鼠标，出现单箭头变成十字双向箭头（✛）时，单击鼠标左键可观察到边界的轮廓及行高度值，如显示为“高度 14.25 （19 像素）”。拖动到适当高度，释放鼠标即可改变行高。以类似方式可改变列宽。

3. 整个工作表的选取

用鼠标左键点击工作表格区的最左上角（即列表 A 的左边、行号 1 的上边），便可选取整张工作表。

若选取了某（些）单元格区域、整行、整列或整张工作表，便可对选取部分进行复制数据、输入数据、修改数据或建立公式等。

1.5 工作表重命名与复制副本

选定某一工作表名称标签，单击鼠标右键会弹出工作表编辑的快捷菜单，选取菜单中的命令，可以分别执行插入、删除、重命名、移动或复制工作表等操作（见图 1-6）。

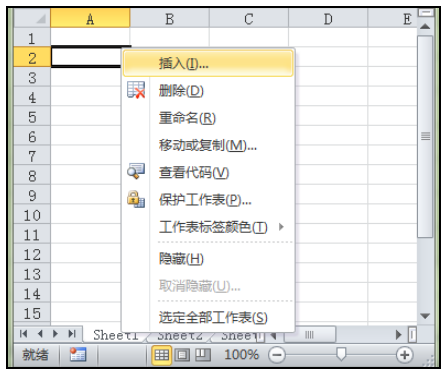


图 1-6 编辑工作表的快捷菜单

1. 工作表名称的颜色与重命名

一般情况下，活动工作表标签背景颜色为白色，非活动工作表标签背景颜色为灰色，为了改变非活动工作表标签背景的颜色，可在标签上单击鼠标右键，在快捷菜单中选取“工

作表标签颜色”（Excel 2003 以上版本新增功能）修改其颜色。对工作表名称进行命名有两种方式，一是在需要重命名的标签上双击，当标签背景变为黑色时，键入文字即可；二是采用如图 1-6 所示的快捷菜单中的“重命名”改变其标签名称。

2. 工作表的插入与删除

单击如图 1-6 所示的快捷菜单中的“插入”，会弹出一个“插入”对话框，可插入常用的工作表、图表等，Excel 2010 插入对话框还有电子方案表格，如贷款分期付款、个人月预算、考勤卡等现存格式表格。如果在插入了一张新的工作表后，想继续插入多个工作表，可按 F4 键进行重复插入。单击如图 1-6 所示的快捷菜单中的“删除”，可删除一张当前工作表。

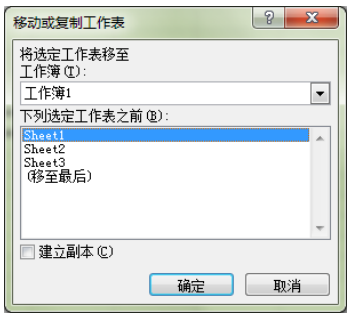


图 1-7 “移动或复制工作表”对话框

3. 工作表的移动与复制

为了在一份工作簿内移动工作表到适当位置，在选取需要移动的工作表后，按住鼠标左键，采用拖放方式，可随意移动工作表。

为了移动或复制工作表到当前工作簿或其他工作簿，可单击快捷菜单中的“移动或复制工作表”，在弹出的对话框（见图 1-7）中，选取某一工作簿、所需移动或复制的位置，或选取“建立副本”，完成移动与复制。

1.6 工作表的格式化

工作表的格式化就是设置工作表的格式，设置格式并不影响其中的数据，设置的主要目的是美化工作表。采用 Excel 编制统计表格时，常需对工作表格式进行设置。

“开始”选项卡中“字体”、“对齐方式”、“样式”等命令组中已显示了部分“格式工具”，一般情况下可采用这些“格式工具”美化工作表内容，如选取单元格和单元格区域后，可对选取部分进行字体、字形、字号、文本颜色、对齐方式等命令修改。

单元格格式设置的另一种方法是：选取单元格和单元格区域，单击鼠标右键，在弹出的快捷菜单中选择“设置单元格格式”命令，在“单元格格式”对话框中可分别对单元格区域的数据类型、对齐方式、字体（特殊效果）、边框线条与颜色、单元格底纹与图案、单元格的锁定与隐含等进行设置。数据类型的设置是数据输入与整理的重要组成部分，将放在第 3 章介绍，下面介绍设置单元格格式的其他方法。

1. 对齐方式

选定要合并的相邻单元格，在图 1-8 中选取“合并单元格”复选框，可将跨越几行或/几列的相邻单元格合并为一个大的单元格。如果单元格区域有多个数据，合并后的单元格内容将只把选定区域最左上角的数据放到合并后所得的大单元格中。

图 1-8 的左上方“文本对齐方式”栏可用于设定水平与垂直的对齐方式。修改图 1-8 中的右上方“方向”，还可改变文本的排列形式。

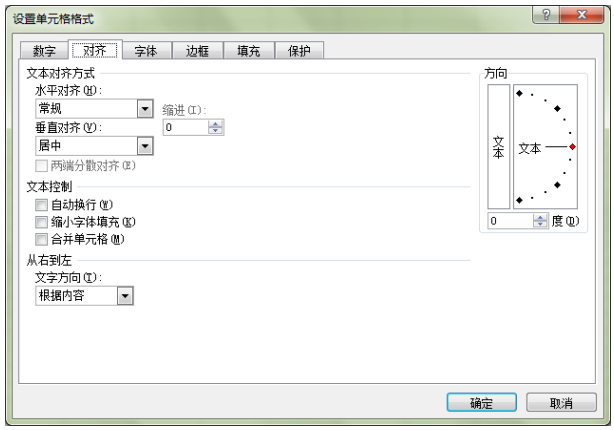


图 1-8　设置对齐方式

2. 字体的特殊效果

有时需要使用字体的上标、下标、删除线（如 5^2、x_1、~~abcd~~）等特殊效果，可先选取需要改变的文本，在图 1-9 中选取相应的复选框，达到预想效果。

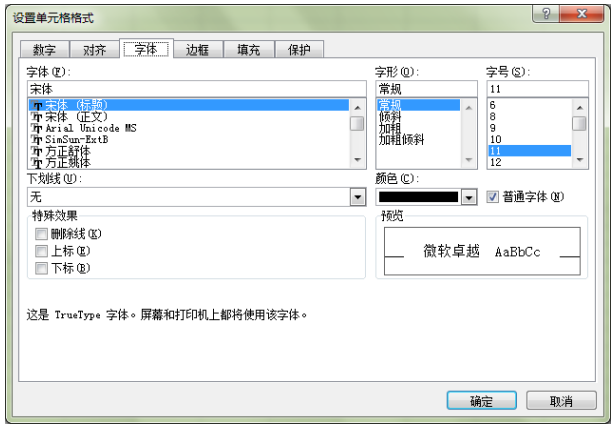


图 1-9　设置字体单元格格式

此处也可设置字体、字形、字号、字体颜色等。

3. 边框线条的设置

要设置单元格的边框，可采用边框选项板（图 1-10 左侧）与边框选项卡（图 1-10 右侧）两种方式。

边框选项板可以快速设置边框，具体操作是：单击“开始→字体→下划线▼”打开边框选项板（见图 1-10 左侧），选取所需边框样式。

边框选项卡可以更详细地设置边框，具体操作是：选定要设置边框的单元格区域，单击右键选取快捷菜单中的“设置单元格格式→边框”（见图 1-10 右侧），先在右侧选取所需的线条样式、颜色，再按图中边框样式设置边框，设置完后单击“确定”。

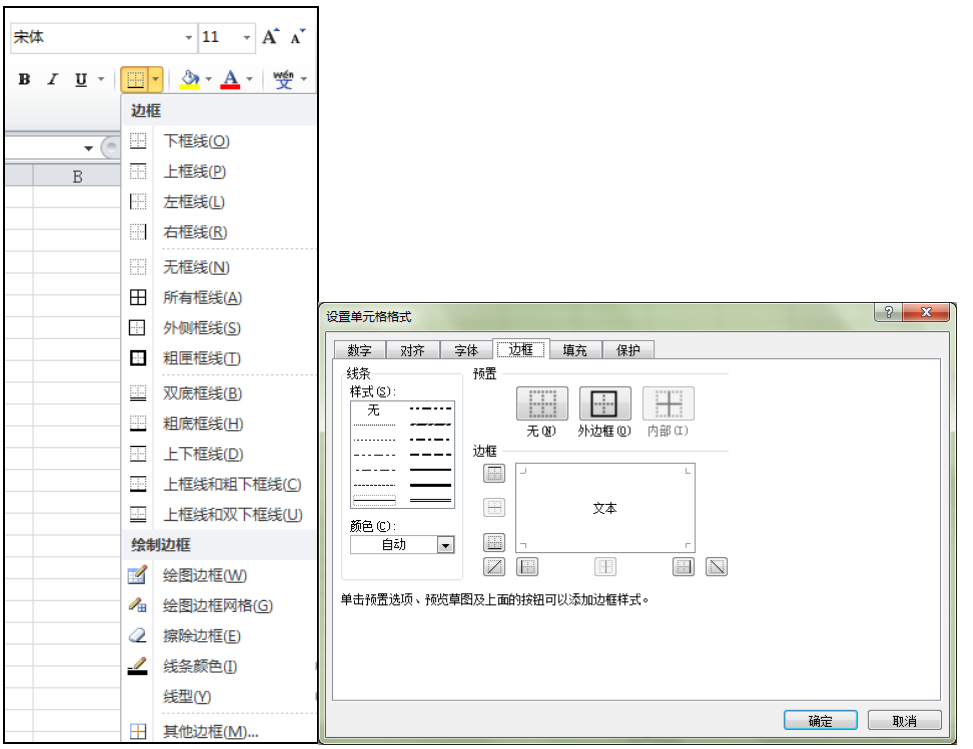


图 1-10　边框选项板（左）与边框选项卡（右）

4. 设置单元格背景图案

要设置单元格的背景颜色，可采用“开始”选项卡的“字体”组中的填充颜色图标 或填充选项卡中的“图案颜色”下拉列表框（见图 1-11）两种方式。

单击填充颜色图标 右侧的下拉按钮可以快速设置单元格背景颜色。

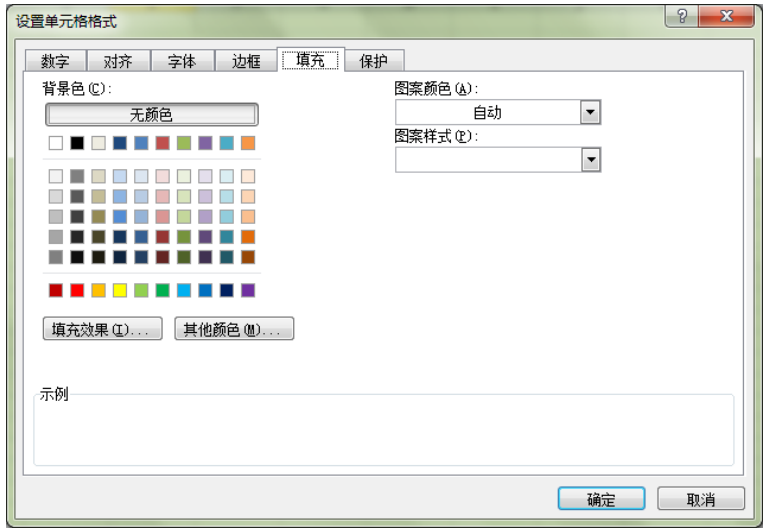


图 1-11 设置单元格图案填充颜色

填充选项卡不仅可以设置图案颜色，而且可以设置背景图案。具体操作是：选定要设置图案的单元格区域，单击鼠标右键选取快捷菜单中的“设置单元格格式→填充→图案样式”（见图 1-11），设置完成后单击“确定”按钮。

5. 格式刷与自动套用格式

快速格式化设置的工具包括格式刷、样式和自动套用格式。

（1）格式刷的使用

具体操作如下。

选定已经包括所需格式的单元格区域；单击常用“开始→剪切板→格式刷”，当将鼠标指针移到工作表中时，鼠标指针变为 形；按住鼠标左键，选定要设置新格式的单元格区域，当鼠标指针扫过这些区域时，它们就自动设置为所需的格式。松开鼠标左键完成设置。

（2）创建样式

用户既可以根据需要将现有单元格格式创建为一种样式，也可以创建为一种全新的样式。根据需要创建一种样式更简单，具体操作步骤如下。

选定已经包括所需格式的单元格区域；然后单击“开始→样式→单元格样式▼→新建单元格样式”，即弹出“样式”对话框；在“样式名”文本框中输入新样式的名称，单击“确定”。

（3）自动套用格式

自动套用格式（见图 1-12）是微软公司经过精心设计的十余种内置格式设置集合，用

户通过套用这些表格，可以迅速格式化工作表中的单元格区域（如统计表），大大提高工作效率。编制统计表时自动套用格式使用频繁。

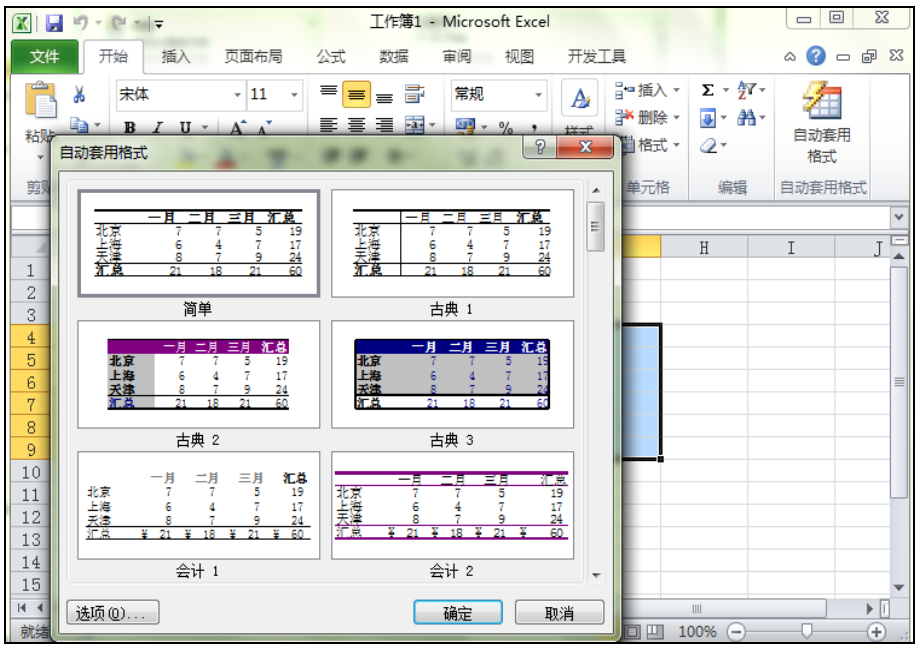


图 1-12 “自动套用格式”对话框

在 Excel 2010 中自动套用格式不像 Excel 2003 那么明显，需要先采用“自定义功能区”设定，以后方可简单调用。设定方法如下。

①单击“文件→选项→自定义功能区”。

②在自定义功能 B 区，即在右侧单击“开始”选项卡，按“新建组”→“重命名”键，在弹出的图形下方“显示名称”右侧框中写入“自动套用格式”，单击“确定”。

③在自定义功能区的左中上部将“常见命令”修改为“所有命令”，在其下方找到“自动套用格式…”。

④在两个功能区的中部单击“添加”按钮，之后单击“确定”按钮。具体设定界面（如图 1-13）所示。

“自动套用格式”图标便在“开始”选项卡功能区的最右侧（见图 1-12）出现。为了利用自动套用格式快速获得表格，可以先选定要应用自动套用格式的单元格区域，单击“自动套用格式”图标，会弹出类似图 1-12 的对话框，从该对话框的格式列表中选择所需的格式，如统计表常用的“简单”格式。

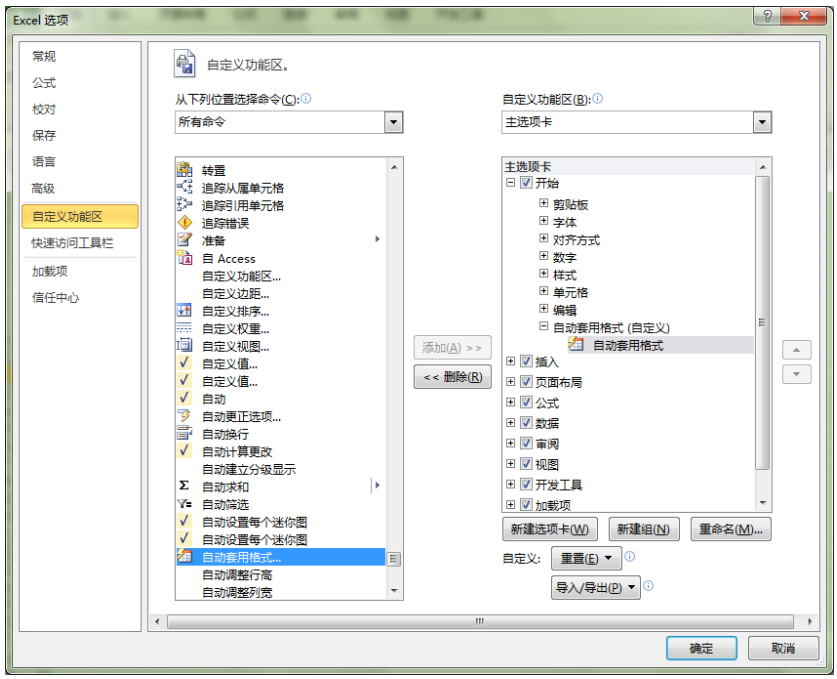


图 1-13　自动套用格式的设定界面

1.7　工作表的打印

建立好工作表之后，用户可以进行页面设置、使用分页符、打印预览、设置打印范围等操作。

1. 页面设置

选择“文件→打印→设置→页面设置”，或点击“页面布局”选项卡的“页面设置”右下方的右下箭头，可弹出图 1-14 的界面。“页面设置”对话框有 4 个选项卡，分别为页面、页边距、页眉/页脚、工作表。采用“页面”选项卡可对页面的打印方向、缩放尺寸、纸张大小、打印质量、起始页码等进行设置。

在“工作表”选项卡中，可以设置打印区域、打印标题、打印顺序等，如果想打印出网格线、行号列标等也可在“工作表”选项卡界面中进行设置（见图 1-15）。“打印标题”也是非常有用的选项，如果需要在每一页都显示标题，那么在“顶端标题行”选取这一标题即可。

如果用户处理的是图形，那么“工作表”选项卡会变成“图表”选项卡，可对图形打印大小、打印质量进行设置。

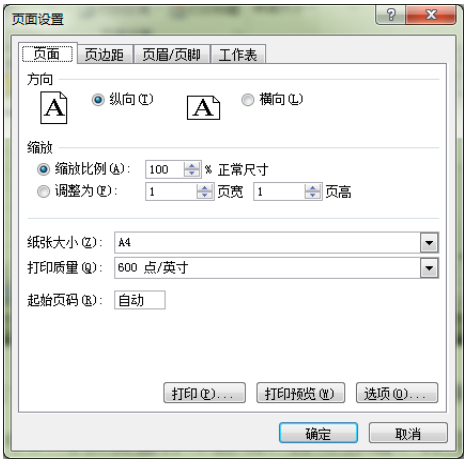


图 1-14 “页面设置”对话框的“页面”选项卡

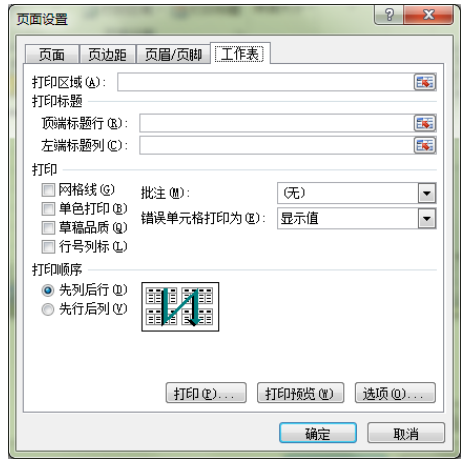


图 1-15 “页面设置”对话框的“工作表”选项卡

2. 使用分页符

单击需要插入分页符之下的行号，或选定该行最左边的单元格，单击“页面布局”选项卡“页面设置”组中“分隔符”命令，然后选择弹出菜单中的“插入分页符”命令即可插入水平分页符。

单击需要插入分页符之右的列标，或选定该列最顶端的单元格，单击“页面布局”选项卡“页面设置”组中“分隔符”命令，然后选择弹出菜单中的“插入分页符”命令即可插入垂直分页符。单击“视图”选项卡“工作簿视图”组中“分页预览”命令，当鼠标指针变成双向箭头时，按住鼠标左键，可将分页符拖到新位置；再选择“工作簿视图”组中“普通”命令，即可退出分页预览视图。

在分页预览视图中，右击垂直分页符右侧或水平分页符下方的单元格，从弹出的快捷菜单中选择“删除分页符”命令，可删除分页符。

单击行号与列标交叉处选定整个工作表，单击“页面布局”选项卡“页面设置”组中“分隔符”，在弹出菜单中选择“重置所有分页符”命令即可删除所有人工设置的分页符。

打印预览和打印范围的设置与一般 Word 文件的打印设置相同。

第 2 章

公式与函数

公式是 Excel 中的精髓和核心，它由单元格内的一系列数值（或者能够代表数值的单元格、单元格区域名称）、运算符、工作表函数组成。利用公式，用户不仅可以进行数学运算（如加、减、乘、除等）、逻辑判断（如真与假），还可以完成工作表数据比较、文本合并等。函数是 Excel 中内置的公式，函数使得各种运算变得更加简便，如果能熟练掌握函数，并能加以利用，则可大大提高工作效率。

2.1 公式的建立

为了建立公式，首先必须键入“=”，键入“=”表示编辑公式即将开始。可以在编辑栏中键入“=”，也可以直接在活动单元格中键入“=”。如果要利用 Excel 函数建立公式，可在编辑栏左侧单击插入函数“ *fx* ”按钮。当然，如果能记住函数名可直接键入函数名及括号。如果想进行求和计算，可在编辑栏（见图 2-1a）或活动单元格 A1（见图 2-1b）中键入“=sum(”。此时，Excel 2003 以上版本会自动提示函数所需的参数。图 2-1 提示：在求和符号的括号中应键入数值 1，数值 2，…，或者键入能代表“数值”的单元格或单元格区域名称。

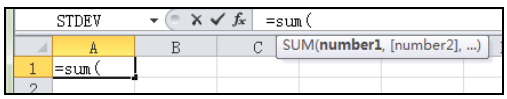


图 2-1a　编辑栏输入函数

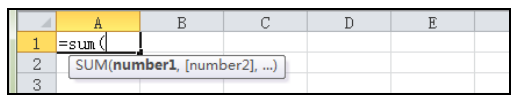


图 2-1b　单元格输入函数

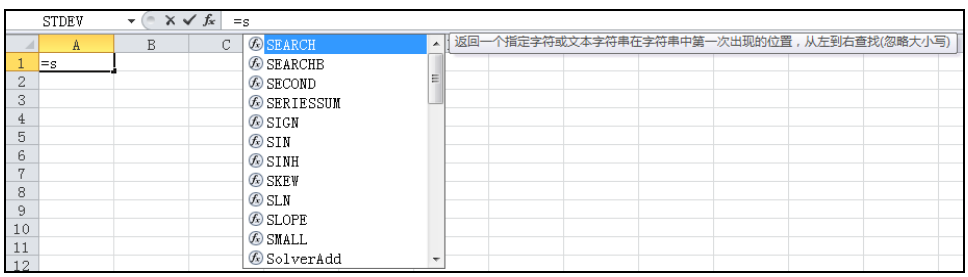


图 2-1c　编辑栏输入函数少数几个字母时的新界面

修改后的 Excel 2010 函数的键入更加人性化，当在编辑栏键入“=”和第一个英文字母后，在其下方会将该英文字母所对应的所有函数均显示出来，供用户选择，且在右侧用文字框，对有关情况进行说明，大大方便了用户进行相关函数的选取，详见图 2-1c 所示。

为了完成公式的编辑，确认输入的公式，可按“Enter”键或用鼠标单击编辑栏左侧的“ √ ”按钮；如果要取消编辑的公式，可单击编辑栏的“×”按钮。

矩阵运算产生的结果是一组数值，而不是一个值，即需要在多个单元格（即一个单元格区域）中建立一个公式，为了完成公式的编辑，确认输入的公式必须按“Ctrl + Shift + Enter”组合键，而不能只按“Enter”键。

另外，在 Excel 2010 的功能区选项菜单中专门设立了“公式”选项卡，除了更加便利醒目的函数库命令组外，还包括了“定义的名称”、“公式审核”、“计算”几个命令组，详见图 2-1d 所示。



图 2-1d　Excel 2010 新有的“公式”选项卡界面

2.2　函数的结构

与公式相比，函数具有特定的语法结构，必须严格遵守，否则 Excel 会弹出错误提示信息。函数以函数名称开始，如 SUM，可大写也可小写（sum）或大小写混合（Sum）；紧接着是左圆括号、参数 1、参数 2、…、右圆括号，当有多个参数时，参数间以逗号分隔。使用函数时，函数名称前的“=”和其后的括号“()”是必不可少的，因此“=”和“()”被称为函数的结构，如图 2-2 所示。

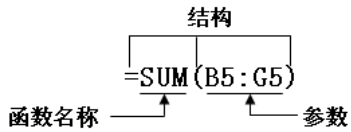


图 2-2　函数的结构

所谓参数是指函数中用来执行操作或计算的值。一个函数可以没有参数，如圆周率函数“PI()”、随机函数“RAND()”等；也可以有一个或多个参数。常用的参数类型有：数值、文本、形如 TRUE 或 FALSE 的逻辑值、数组、引用单元格（见图 2-2）、错误值 #N/A 等，常量、公式或其他函数也可作为参数出现在函数中。比如在 IF 函数“=IF(AVERAGE(B3:G3)>50,SUM(B4:G4),0)”中，有 AVERAGE 函数、SUM 函数作为其参数。



注意：

①给定的参数必须能够产生有效的值；

②当函数或公式中没有可用数值时，将产生错误值 #N/A，公式在引用错误值 #N/A 时，将不进行数值计算，而是返回 #N/A；

③常量是指直接键入到单元格或公式中的数字或文本，如日期 2007-11-19、数值“100”、文本“西雅图”等都是常量，公式本身或公式得出的值不是常量。

2.3 公式中的运算符

在 Excel 中有 4 种类型的运算符，即算术运算符、比较运算符、文本运算符、引用运算符。

（1）算术运算符

算术运算符可以完成基本的数学运算，如加、减、乘、除、乘方（如乘幂的 2 次方）、开方（如乘幂的 1/2＝0.5 次方）等（见表 2-1）。

表 2-1　算术运算符

运 算 符	运算符名称	含　义	操作举例	运算结果
+	加号（plus）	加（addition）	=10+5	15
-	减号（minus）	减（subtraction）	=10-5	5
*	星号（asterisk）	乘（multiplication）	=10*5	50
/	斜杠（slash）	除（division）	=10/5	2
^	脱字符（caret）	乘幂（exponentiation）	=10^5	100000
%	百分号（percent sign）	百分比（percentage）	=200%	2

一般的计算机程序中没有开方的运算符，开方需要采用乘幂运算符完成。对于常数 a 开 n 次方，即计算 $\sqrt[n]{a}$，Excel 中可在编辑栏输入公式“=a^(1/n)”完成运算。如要对 100000 开 5 次方（即 $\sqrt[5]{100000}$ ），可采用“=100000 ^(1/5)”或“=100000 ^ 0.2”进行运算（结果=5）。2^{-2} 可采用“=2 ^(-2)”进行运算（结果=0.25）。其他算术运算符及其操作举例见表 2-1。

（2）比较运算符

比较运算符的种类及其含义见表 2-2。

表 2-2　比较运算符

运 算 符	运算符名称	含　义	操作举例
=	等于号	等于	A1=B1
>	大于号	大于	A1>B1
<	小于号	小于	A1<B1
>=	大于等于号	大于等于	A1>=B1
<=	小于等于号	小于或等于	A1<=B1
<>	不等于号	不等于	A1<>B1

（3）文本运算符

在 Excel 中，文本运算符只有一种，即“&”，该运算符可将文本或数值连接成串。文本运算符的举例见表 2-3。

表 2-3　文本运算符

运 算 符	运算符名称	含　义	操作举例	结　果
&	连字符	连接数据或文本	="*t*" & "检验"	*t* 检验
			=A1&B1（注：单元格 A1 为数字 2013，B1 为数字 125）	2013125

注意：对于较低的 Excel 版本，如果是连接数字，产生的结果表面上是数字，但实际上它的格式是文本，如果需要利用产生的结果“数字”参与计算，必须选取数字所在的区域，通过“设置单元格格式”将其格式改变为数值（也可通过复制与选择性粘贴数值来改变其格式）。也可对文本“数字”进行乘法运算（如用“*1”，即乘以 1），改变文本格式“数字”成数值格式“数字”。

（4）引用运算符

除了以上运算符外，当 Excel 公式引用一个单元格区域或多个单元格区域时，还需要采用引用符号，常用的引用运算符有“:”（冒号）、“,”（逗号）、“ ”（空格），分别表示引用范围、合并引用、交叉引用。

当引用一个单元格区域时，需要知道区域的左上角位置和右下角位置，如左上角位置为第 2 列第 5 行（B5），右下角位置为第 4 列第 8 行（D8），则该区域可表示为“B5:D8”（引用范围为 B5:D8）。如果需要计算三个单元格区域“B5:D8”、“A1:A6”、“H1:H5”之间的和，可以在编辑栏输入“=SUM(B5:D8, A1:A6, H1:H5)”（即求 B5:D8、A1:A6、H1:H5 三个单元格区域的和）来完成。

合并引用 SUM(C5:C10, A6:E8)与**交叉引用** SUM(C5:C10, A6:E8)的计算结果完全不同。就图 2-3（由单击“公式→公式审核→显示公式”获得此界面）的（左侧）数据而言，合并引用 SUM(C5:C10, A6:E8)的含义是先将 C5:C10 数据相加（即 6 个 3 相加=18），再将 A6:E8 数据相加（即 3 个 1、2、3、4、5 相加=45），最后结果是 63；交叉引用 SUM(C5:C10A6:E8)的含义是单元格区域 C5:C10 与 A6:E8 相交部分的数据相加（即 3 个 3 相加=9），最后结果是 9。想回到数据结果界面，再次单击“公式→公式审核→显示公式”即可。

	A	B	C	D	E	F	G
1	1	2	3	4	5		
2	1	2	3	4	5		
3	1	2	3	4	5		
4	1	2	3	4	5		
5	1	2	3	4	5		
6	1	2	3	4	5		=SUM(C5:C10 A6:E8)
7	1	2	3	4	5		=SUM(C5:C10,A6:E8)
8	1	2	3	4	5		
9	1	2	3	4	5		
10	1	2	3	4	5		

图 2-3　合并引用与交叉引用的差别

2.4 运算符的优先顺序

如果公式中包含了相同优先级的运算符（如+、–），则从左到右进行计算。当公式中包括多个运算符时，往往出现优先级别不同的情况。只有弄清了运算符的优先级别，哪些运算符的优先级别相同，哪些运算符的优先级别排在前面，哪些排在后面，才不至于造成运算错误。下面的表 2-4 列出了 Excel 运算符优先级别的顺序。

表 2-4　Excel公式中运算符的优先级顺序

优先级顺序	运 算 符	说　　明
1（高）	:（冒号）	引用运算符
2	（空格）	
3	,（逗号）	
4	-	负号
5	%	百分比
6	^	乘幂
7	*和/	乘和除
8	+和-	加和减
9	&	连接两串文本（连接符）
10（低）	=、>、<、>=、<=、<>	比较运算符

在表 2-4 中，由上至下运算符优先级别由高到低。例如有“*”（乘）和“+”（加）同时存在时，如果没有括号限制，不管其先后排列顺序如何，均应先算乘法、后算加法。

括号“()”可以改变计算的优先级别，先计算括号内再计算括号外，有多个括号时先计算里层的括号，再计算外层的括号。在不知道运算符的优先级别时，不妨使用括号。



注意：每个左括号必须有一个对应的右括号，尽管 Excel 2003 以后版本可以自动补足所缺括号，但编辑公式时也应谨防出错。如果左右括号的数目不对应，Excel 会将每个括号以不同的颜色显示，以便找出错误。

2.5 单元格的混合引用

所谓单元格引用就是在公式中标识工作表数据单元格或单元格区域所在的位置，通过引用可在同一个公式中使用工作表不同位置的数据，或者在多个公式中使用同一个单元格的数值，此外，还可引用同一工作簿不同工作表的单元格等。

公式中如果需要引用第 2 列第 5 行的单元格内容，一般采用 B5 表示，这种引用叫做相对引用，其引用的公式改变位置时，单元格引用位置也随之发生相应改变；为了使引用的公式改变位置时，而其引用单元格位置不发生改变，需要采用绝对引用，即在列标字母和行号数字之前分别加上美元符号（$），如$B$5；除此之外，还有一种引用叫做混合引用，混合引用有两种表示方法，分别为$B5 和 B$5。以上 4 种引用的表示方式不同，其实际含义也不一样。区分与灵活运用单元格的相对引用、绝对引用与混合引用在编辑公式时十分重要，这样可简化许多数据处理的运算问题。

2.5.1　相对引用

相对引用就是使用单元格所在的列标和行号作为引用，如引用 B5，这是最常用的单元格引用方法。

相对引用有这样的特点：将相应计算公式复制或填充到其他单元格时，其中的单元格引用会自动随着公式移动的位置发生相对变化。例如，将 B6 单元格的公式“=SUM(B1:B5)”（求 B1、B2、B3、B4、B5 单元格的值之和，即等价于“= B1+B2+B3+B4+B5”）复制到 C6、D6、E6，其公式内容会相应地变为“=SUM(C1:C5)”、“=SUM(D1:D5)”、“=SUM(E1:E5)”。由于其位置向右移动了 1、2、3 列，其列标也相应地由 B 变为 C、D、E。

这样，当用户需要使用大量类似的公式时，可先输入一个公式，然后将公式复制或拖放填充到其他的相应单元格即可，从而节约了大量输入公式时间。

2.5.2　绝对引用

有时，在公式中需要始终引用一个固定单元格的内容。例如，计算率比时，计算公式的分母是固定的。无论将公式复制或填充到什么位置都不希望公式中的分母发生改变，此时就需要使用单元格的绝对引用。绝对引用的方法是在列标和行号前分别加上“$”符号，例如，$D$6 表示工作表单元格 D6 的绝对引用，而$B$1:$E$6 则表示单元格区域 B1:E6 的绝对引用。

例如，单元格 A1 至 A5 分别为某年某地猩红热、麻疹、百日咳、白喉、痢疾等 5 种疾病的发病人数，单元格 A6 为合计病人数，为了计算各种疾病的构成比，在 B1 单元格建立计算公式为“=A1/$A$6”，将此 B1 单元格公式复制或拖放到单元格 B2、B3、B4、B5 中，公式分别自动改变为“=A2/$A$6”、“=A3/$A$6”、“=A4/$A$6”、“=A5/$A$6”，达到了计算率比的目的。

2.5.3　混合引用

有时，希望单元格引用的公式，随着复制粘贴或拖放的单元格位置横向移动或纵向移

动，具有不同的变化规律，这时需要采用混合引用。

混合引用有两种情况：

① 单元格引用公式复制、粘贴或拖放横向不变，而纵向改变，如$B2（见图 2-4 左）。这是一种列标不变、行号改变，即列（标）绝对引用，行（号）相对引用的情况；

② 单元格引用公式复制、粘贴或拖放横向改变，而纵向不变，如 B$2（见图 2-4 右）。这是一种列标改变、行号不变，即列（标）相对引用，行（号）绝对引用的情况。



图 2-4　混合引用的两种情况

2.5.4　A1 引用样式

Excel 中有 A1 引用样式与 R1C1 引用样式之分，本书只采用 Excel 默认的 A1 样式。以上所给的绝对引用、相对引用、混合引用例子都是 A1 引用样式。表 2-5 比较全面地反映了 A1 引用样式及其相应含义。

表 2-5　A1 引用样式及其含义

样　式	含　义
A5	第 A 列第 5 行对应的单元格
A5:A20	第 A 列第 5 行至第 20 行对应的单元格区域（共有 16 个纵向单元格）
A5:E5	第 A 列第 5 行至第 E 列第 5 行对应的单元格区域（共有 5 个横向单元格）
5:5	第 5 行中的全部单元格
5:20	第 5 行至第 20 行中的全部单元格
A:A	第 A 列中的全部单元格
A:E	第 A 列至第 E 列中的全部单元格
A5:E20	第 A 列第 5 行至第 E 列第 20 行对应的单元格区域（共有 5×16 个单元格）

此外还有三维引用，即工作簿中多个工作表的单元格引用，如在 Sheet6 工作表中求 Sheet1 至 Sheet5 工作表的单元格区域 A5:E20 的总和，可选取 Sheet6 工作表中某一单元格，在编辑栏中输入公式“=SUM(Sheet1:Sheet5!A5:E20)”，按“Enter”键结束公式的输入。

2.5.5　应用实例

相对引用的公式所在单元格与所引用单元格是一种相对关系；绝对引用的公式与所引用单元格是一种绝对关系；而混合引用两者兼而有之。绝对引用与相对引用的主要差异在

于：将含有公式的单元格复制或填充到其他单元格时，公式中绝对引用不变，而相对引用随单元格位置的变化而相应发生改变。用户可根据所要完成的任务，合理使用相对引用、绝对引用或混合引用。

（1）九九乘法口诀表的编制

乘法口诀表有 $9\times9=81$ 格，如果每一格都输入一个公式，则需输入 81 个公式。下面采用混合引用只需建立一个公式，然后由该公式复制、粘贴或填充到其他 80 个单元格来完成乘法口诀表的编制。

如果采用相对引用，在 B2 单元格的计算 $1\times1=1$ 的公式应该是“A2*B1”。其他单元格的计算公式也都是由第 A 列数字 1～9 的其中一个数，乘以第 1 行数字 1～9 的其中一个数。引用第 A 列时，列标不变、行号改变，因此采用混合引用应将单元格 B2 中公式的“A2”修改为“$A2”；引用第 1 行时，列标改变、行号不变，因此采用混合引用应将单元格 B2 中公式的“B1”修改为“B$1”。这样 B2 单元格中的计算公式修改为“$A2*B$1”，将该公式复制、粘贴或填充到其他 80 格，即可获得九九乘法口诀表（见图 2-5，也见光盘中“02-01 乘法口诀表.xlsx”文件）的编制。

	A	B	C	D	E	F	G	H	I	J
1		1	2	3	4	5	6	7	8	9
2	1	1	2	3	4	5	6	7	8	9
3	2	2	4	6	8	10	12	14	16	18
4	3	3	6	9	12	15	18	21	24	27
5	4	4	8	12	16	20	24	28	32	36
6	5	5	10	15	20	25	30	35	40	45
7	6	6	12	18	24	30	36	42	48	54
8	7	7	14	21	28	35	42	49	56	63
9	8	8	16	24	32	40	48	56	64	72
10	9	9	18	27	36	45	54	63	72	81

图 2-5　九九乘法口诀表的编制

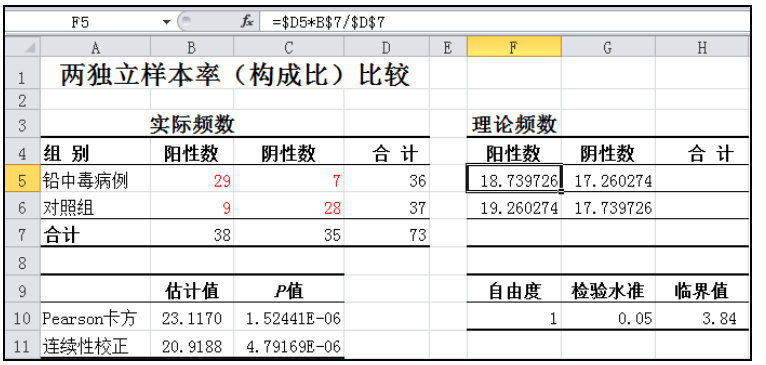
填充方法更简洁，其方法是：将鼠标移至 B2 单元格右下角的填充柄，此时空心十字光标变为实心十字光标，按下鼠标左键，向右拖放至单元格 J2，然后在单元格 J2 右下角的填充柄处再次按下鼠标左键，向下拖放至单元格 J10。

（2）四格表卡方检验理论频数的计算

四格表卡方检验的某格子理论频数等于该格子对应的行合计与列合计之积与所有格子观察个体总数的比，即 $T_{RC}=\dfrac{n_R n_C}{n}$。在单元格区域 A3:D7 处列出了四格表资料，B5:C6 为四格表的实际频数所在单元格位置，D5:D6 为行合计，B7:C7 为列合计，D7 为所有观察个体总数。拟将计算的理论频数放在 F5:G6。

如果仅采用相对引用建立公式，则计算 4 个理论频数需要采用 4 个公式，F5、G5、F6、G6 单元格公式分别为“=D5*B7/D7”、“=D5*C7/D7”、“=D6*B7/D7”、“=D6*C7/D7”。如果在单元格 F5 中建立公式“= $D5*B$7/$D$7”，然后将此拖放到 G6 即可快速完成卡方

检验理论频数的计算（见图 2-6，也见光盘中“02-02 四格表卡方检验理论频数.xlsx”文件）。



F5 =$D5*B$7/$D$7

	A	B	C	D	E	F	G	H
1	两独立样本率（构成比）比较							
2								
3		实际频数				理论频数		
4	组 别	阳性数	阴性数	合 计		阳性数	阴性数	合 计
5	铅中毒病例	29	7	36		18.739726	17.260274	
6	对照组	9	28	37		19.260274	17.739726	
7	合计	38	35	73				
8								
9		估计值	P值			自由度	检验水准	临界值
10	Pearson卡方	23.1170	1.52441E-06			1	0.05	3.84
11	连续性校正	20.9188	4.79169E-06					

图 2-6　四格表卡方检验理论频数的计算

2.6　编辑公式常用的快捷键

（1）F4 功能键

编辑公式时，单元格绝对引用和混合引用需要输入“$”号，绝对引用要分别在列标和行号之前输入“$”号，混合引用要在列标或行号之前输入“$”号。输入“$”号时，需要将输入方式改变为英文，且须按 Shift 键输入“$”号。为了操作更简便，可采用 F4 功能键解决此问题。如需对单元格 A1 加“$”号，将光标放在 A1 的旁边，然后第一次按 F4 功能键 A1 会变为$A$1，第二次按 F4 功能键会变为 A$1，第三次按 F4 功能键会变为$A1，第四次按 F4 功能键又会返回到相对引用状态，即 A1，……。

（2）Ctrl+`键

这里“`”撇号位于键盘的左上方，数字键“1”的左边。

在默认状态下，在输入公式的单元格中显示的只是结果，而不是公式。若想在单元格中显示公式，可以使用快捷键“Ctrl＋`”进行公式内容与结果的切换。

（3）Ctrl+‘键

将当前单元格上方单元格中的公式复制到当前单元格或编辑栏，公式不变。“‘”为单引号，位于大键盘中偏右。

2.7　编辑公式时常见的错误信息

以下介绍 8 种常见的错误信息出现原因及其解决办法。

（1）########

①输入到单元格中的数据太长、单元格公式所产生的结果太大、日期显示长度超过单元格宽度，在单元格中显示不下时，将在单元格中显示########。可以通过调整列标之间的边界来修改列的宽度。

注意，对于第二代身份证含有 18 位数字（如 420104201301271867），在一般情况下，输入身份证号码后，不是显示为“########”，而是显示为科学计数法数字（如 4.2E+17），此时的数字是以单元格“常规格式”显示，如果修改为该单元格数字格式为不包含小数的“数值格式”，则该身份证号变为“420104201301271000”，即后面三位变为 0（原因是 Excel 最多只能保留 15 位有效数字）。在 Excel 中正确输入身份证号码的办法是：事先设定数字的单元格格式为“文本”格式；或先输入“'”（英文的单引号，然后再输入 18 位数字，此时后面的数字将以文本方式保存。

②对日期或时间做减法时出现负值。Excel 中的日期和时间必须为正值。如果日期或时间产生了负值，将在整个单元格中显示########。如果仍要显示这个数值，须单击“开始→数字”选项，然后单击“数字”命令组右下角的 符号，然后选定一个不是日期或时间的格式。

（2）#DIV/0!

①输入的公式中包含明显的除数 0，例如“=−120/0”，则会产生错误信息 DIV/0!。

②在公式中除数使用了空单元格（当运算对象是空白单元格时，Excel 将此空值解释为零值）或包含零值单元格的单元格引用。

解决办法是：修改单元格引用，或者在用作除数的单元格中输入不为零的值。

（3）#VALUE!

①在需要赋单一数据的运算符或函数时，却赋给了一个数值区域。

②当使用不正确的参数或运算符时，或者当执行自动更正公式功能时不能更正公式。

③在需要数字或逻辑值时却输入了文本，Excel 不能将文本转换为正确的数据类型。

这时应确认公式或函数所需的运算符或参数正确，并且公式引用的单元格中包含有效的数值。例如，单元格 B3 中有一个数字，而单元格 B4 包含文本，则公式“=B3＋B4”将返回错误信息＃VALUE!。

（4）＃NAME?

在公式中使用了 Excel 所不能识别的文本时将产生错误信息。可以从以下几方面进行检查、纠正错误。

①如果使用了不存在的名称而产生这类错误，应确认使用的名称确实存在。在功能区中单击“公式→定义的名称→定义名称/名称管理器”来确认，如果所需名称没有被列出，使用“定义”命令添加相应的名称；

②如果名称正确，但函数名拼写有错误，则应修改函数名的拼写错误；

③确认公式中使用的所有区域引用是否使用了冒号（:）。例如：SUM(A2:C20)。

注意：公式中的文本应放在双引号中。即如果公式中使用到了文本，那么要将文本放在双引号中。

（5）#N/A

在函数或公式中没有可用数值时产生的错误信息。如果某些单元格暂时没有数值，可以在这些单元格中输入“#N/A”。这样，公式在引用这些单元格时不进行数值计算，而是返回“#N/A”。

（6）#REF!

单元格中出现这样的信息是因为该单元格引用无效的结果。例如，删除了有其他公式引用的单元格。

（7）#NUM!

这是在公式或函数中某个数字有问题时产生的错误信息。例如，在需要数字参数的函数中使用了不能接受的参数，或者公式产生的数字太大或太小等。

（8）#NULL!

单元格中出现此错误信息是试图为两个并不相交的区域指定交叉点时产生的错误。例如，使用了不正确的区域运算符或不正确的单元格引用等。

其实，为了尽快找到公式中出现的错误，我们还可以按以下步骤进行检查。

①看所有圆括号是否都成对出现；

②检查是否已经输入了所有必选的参数；

③在函数中输入函数嵌套时是否超过了级数；

④引用的工作簿或工作表名称中包含非字母字符时是否使用了单引号；

⑤每一个外部引用包含的工作簿的名称及路径是否正确；

⑥在公式中输入数字时是否为它们设置了格式等。

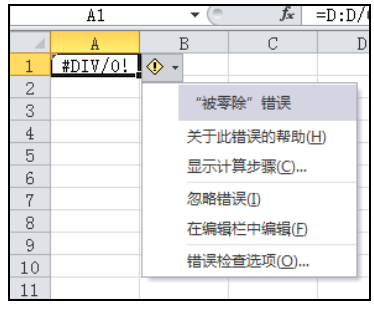


图 2-7　有关错误信息的快捷菜单

Excel 2003 以上版本在以上几种错误信息提示符的左上角标注了绿色小三角，当选取“标注”单元格后，单元格左侧会出现一个“!”，鼠标移至“!”，单击其右侧下三角，会弹出有关错误信息的一个菜单（见图 2-7），使得查找错误与修改错误更加快捷、容易。

此外，Excel 的计算结果有时会令人费解，如计算“=2.02-2.01”，如果保留 16 位小数，得到 0.0100000000000002；计算“=8.02-8.01”，如果

保留 16 位小数，得到 0.0099999999999998。其原因是 Excel 将十进制数值首先转换为二进制进行计算，然后将计算结果再转换回到十进制。

2.8　公式保护与撤销保护的技巧

出于安全或保密的目的，不想让他人修改或看到自己的公式，可以将公式进行锁定或/和隐藏。在默认情况下，整个表格的单元格都是锁定的，但由于工作表没有被保护，因此锁定不起作用。

（1）锁定整个工作表

单击“审阅→更改→保护工作表”，两次重复输入相同密码后，单击“确定”。此时工作表即被保护，单元格锁定生效。

（2）隐藏整个工作表

①将光标移至工作表左上侧的列标行号交界处，单击选取整个工作表；

②单击“开始→单元格→格式▼→设置单元格格式”，选择“保护”选项，此时可见锁定复选框前已打对钩，此外也在隐藏复选框前已打对钩，单击“确定”；

③单击“开始→单元格→格式▼→设置单元格格式”，选择“可见性”选项，单击“隐藏或取消隐藏”→“隐藏工作表”；

④单击“审阅→更改→保护工作表”，两次输入相同密码后，单击“确定”，此时工作表被隐藏。

（3）公式的锁定

很多情况下，只想对公式所在单元格进行保护，而无须对其他单元格进行保护。这种情况下，可先选取整个工作表（将光标移至工作表左上侧的列标行号交界处），单击“开始→单元格→格式▼→设置单元格格式”，选择“保护”选项，去掉锁定复选框前的对钩，单击“确定”；再选取公式所在单元格，单击“开始→单元格→格式▼→设置单元格格式”，选择“保护”选项，再在锁定复选框中打对钩，单击“确定”；单击“审阅→更改→保护工作表”，两次输入相同的密码后，单击“确定”，公式便被锁定。此时，如何其他地方可以正常编辑、修改，而公式所在单元格被保护。

（4）留出存放与修改数据的空间，工作表其他部分被保护

如果希望某一选定的单元格区域（如统计分析数据存放区）可以修改，工作表的其他部分均被锁住，可先选取该单元格区域，单击“开始→单元格→格式▼→设置单元格格式”，选择“保护”选项，去掉锁定复选框前的对钩，单击“确定”；再单击“审阅→保护工作表”，两次输入相同的密码后，单击“确定”即可实现这一愿望。

（5）撤销保护

如果要撤销对工作表的保护，则单击“审阅→撤销保护工作表”，这时会要求使用者输入密码。输入正确的密码后，就可任意修改工作表了。锁定和隐藏的区别在于：仅仅锁定，其公式仍然可以看见，但不能被修改；而仅仅隐藏，可避免他人看到该单元格公式。可同时选择锁定和隐藏，也可单独选择。如果设有密码，应牢记所设密码。

（6）破解被保护工作表密码

当忘记了“保护工作表”的密码，或者需要破解网上下载的 Excel 格式小程序，可按照以下步骤操作：

①打开需要破解保护密码的 Excel 文件（如提示有宏文件，应该启用宏文件）；

②单击“开发工具→录制宏”，输入宏名（如:a），单击“确定”，立即“停止录制”（这样得到一个空宏）；

③单击“宏”，在“宏”对话框选取刚创建的宏文件名“如：a”,单击“编辑”；

④删除窗口中的所有字符（只有几行），复制粘贴光盘文件“破解被保护工作表密码方法.docx”中的“代码”，关闭编辑窗口，返回 Excel 主界面；

⑤单击“开发工具→宏”，在“宏”对话框选取“AllInternalPasswords”（“a”文件名已被“AllInternalPasswords”取代），单击“执行”，然后单击“确定”两次，便可获得 Excel 工作表密码。

如果忘记了公式的保护密码，但又想超越工作表保护，看到以前你设置的公式，可选中设有密码的单元格或工作表，采用“复制”与“粘贴”，将其拷贝到一个新的工作簿（注意：一定要是新工作簿）中即可。

采用快捷键“Ctrl+`”可进行公式内容与结果的切换。但当隐藏公式后，该快捷键将不再起作用。即使是在公式编辑栏中使用快捷键“Ctrl+`”，也看不到公式的内容。

2.9 常用数学函数应用举例

Excel 2003 为用户提供了财务、日期与时间、数学和三角、统计、查询和引用、数据库、文本、逻辑、信息、工程 10 类函数；Excel 2007 在此基础上增加了多维数据集函数；Excel 2010 在日期与时间、数学和三角、统计函数方面又新增了一些函数，这些函数可以在最新版本的 Excel 中使用，但与早期版本的 Excel 不兼容。

Excel 函数是 Excel 功能的最精华部分，对于 Excel 2003、Excel 2007 和 Excel 2010 的所有函数见附录 A 所示。部分与统计有关的函数还将在后面章节中提到。下面只列举几个常用数学函数及其有关应用。

2.9.1　常用数学函数

1. 圆周率

圆周率 π ＝3.14159265358979，可采用 Excel 公式“=PI()”进行计算。PI 后面是括号，括号内是空的，无任何内容。

2. 对数与指数

（1）对数

假使有一个常数 a，以 e＝2.7183 为底的自然对数、以 10 为底的对数、以 2 为底的对数可分别表示为 $\ln a$、$\log_{10}(a)$、$\log_2(a)$。

对于自然对数，Excel 函数为 LN(a)，例如，$\ln 10$ 的 Excel 计算公式为“=LN(10)”，计算获得的结果为 2.3026。

对于其他对数可采用 Excel 函数 LOG（a，基数）计算，公式中 a 为常数，基数即对数的底。例如，计算 $\log_{10} 20 = \lg 20$ 的 Excel 计算公式为“=LOG(20,10)”，计算获得的结果为 1.3010；计算 $\log_2 20$ 的 Excel 计算公式为“=LOG(20,2)”，计算获得的结果为 4.3219。

计算 $\log_{10} 20$ 的另外一种 Excel 计算公式为“=LOG10(20)”。

（2）指数

对于自然对数的指数 e^a，Excel 函数为 EXP(a)，例如，e^5 的 Excel 计算公式为“=EXP(5)”，计算获得的结果为 148.4132。

对于其他指数 x^y，可用乘幂运算符“^”计算。例如，10^5 的 Excel 计算公式为“=10^5”，计算获得的结果为 100000。

开方也可用乘幂运算符“^”计算，例如，将 $\sqrt[3]{10}$ 写为 $10^{1/3}$，Excel 计算公式为“=10^(1/3)”，计算获得的结果为 2.1544。

3. 阶乘

当用四格表精确概率法计算概率时，用到阶乘（factorial）的计算。数学上一个非负整数 n(包括 0)的阶乘计算公式为：$n! = n \times (n-1) \times \cdots \times 2 \times 1$，例如 0！=1；1！=1；5！$= 5 \times 4 \times 3 \times 2 \times 1 = 120$。

阶乘函数为：FACT(数值)，括号内的“数值”表示一个非负的数值。例如，计算 5！的 Excel 公式为“=FACT(5)”，计算获得的结果为 120。

Excel 还有一个双倍阶乘函数：FACTDOUBLE(数值)。

若“数值”为奇数，有：FACTDOUBLE(n)=n(n-2)(n-4)…(3)(1)。如 FACTDOUBLE(5)=$5 \times 3 \times 1$=15，Excel 公式为“=FACTDOUBLE(5)”，计算获得的结果为 15。

若“数值”为偶数，有：FACTDOUBLE(n)=n(n-2)(n-4)…(4)(2)。如 FACTDOUBLE(6)=$6 \times 4 \times 2$=48，Excel 公式为“=FACTDOUBLE(6)”，计算获得的结果为 48。

4. 排列与组合

（1）排列

从 n 个不同对象中取出 k 个对象，并加以有次序排列的方法称排列（permutation），用数学符号可表示为：

$$P_n^k = n(n-1)(n-2)\cdots(n-k+1) = \frac{n!}{(n-k)!}$$

Excel 排列函数为：PERMUT(总数=n，选定数值=k)。例如，从 5 个不同对象中取出 2，Excel 公式为“=PERMUT(5,2)”，计算获得的结果为 20。

（2）组合

从 n 个不同对象中取出 k 个对象，并加以组合（不管其次序）的方法称组合（combination），用数学符号可表示为：

$$\binom{n}{k} = \frac{P_n^k}{k!} = \frac{n!}{k!(n-k)!} = C_n^k$$

Excel 组合函数为：COMBIN(总数=n,选定数值=k)。例如，从 5 个不同对象中取出 2，Excel 公式为“= COMBIN(5,2)”，计算获得的结果为 10。

5. 随机数

为了产生 0～1 之间的随机数，可采用函数：RAND()。类似于 PI()，括号中是空的，不需要加任何成分；获得的随机值是 0～1 之间的小数。例如，某次由“= RAND()”获得 0.2615467，通过公式的复制、粘贴或填充后，随机数改变为 0.2747219、0.8055239、……。

为了获得给定的任意两个数（包括正数、0、负数）之间的随机数，可采用函数：RANDBETWEEN(较小值,较大值)。注意：该函数获得的随机数为整数。例如，某次由“=RANDBETWEEN(-100,100)”获得的随机数为 21，通过公式的复制、粘贴或填充获得的随机数为-44、41、……。

请注意随机数函数所得的结果不像 PI()是常数，而是变数。大多数情况下，每次运行都会得到不同的值。

6. 数据取舍函数

下面介绍 4 个数据取舍函数：

- 符号函数：SIGN(数值)；
- 取整函数：INT(数值)；
- 截断取舍函数：TRUNC(数值,[保留小数点位数])；
- 四舍五入取舍函数：ROUND(数值，保留小数点位数)。

这 4 个函数之间的区别详见图 2-8（也见光盘中“02-03 数据取舍函数.xlsx”文件）。

为了获得数据的符号，可采用符号函数：SIGN(数值)。当数值为负数时，函数计算结

果为-1；数值为正数得 1；数值为 0 得 0。

D7　=TRUNC(A7,0)

	A	B	C	D	E
1	数值	符号函数	取整函数	截断取舍函数	四舍五入函数
2	0.00	0	0.00	0.00	0.00
3	0.50	1	0.00	0.00	0.50
4	-1.25	-1	-2.00	-1.00	-1.30
5	2.00	1	2.00	2.00	2.00
6	2.75	1	2.00	2.00	2.80
7	-3.50	-1	-4.00	-3.00	-3.50
8	4.25	1	4.00	4.00	4.30
9	5.00	1	5.00	5.00	5.00
10	-5.75	-1	-6.00	-5.00	-5.80
11	-6.50	-1	-7.00	-6.00	-6.50
12	7.25	1	7.00	7.00	7.30

图 2-8　4 种数据取舍函数的比较

为了只取数据的整数部分，可采用取整函数：INT(数值)。注意：对于正数，保留了整数部分，小数部分全部舍去；但对于负数，小数部分舍去，但获得的整数部分为原整数加 1。如“=INT(–3.50)”等于–4，而不是等于–3。取整函数得到的结果均小于原数值，无论正数还是负数。

比取整函数更灵活的一个取舍数据函数是截断取舍函数：TRUNC(数值，[保留小数点位数])(这里的“[…]”表示为可选项)。当保留小数点位数为 0 时，获得整数，但与取整函数获得的结果有所不同，负数仍为原整数，如“=TRUNC(–3.5,0)”等于–3。

为了四舍五入取舍，可采用 Excel 函数：ROUND(数值,保留小数点位数)。如“=ROUND(–1.25,1)”等于–1.3。

此外，Excel 还有很多其他取舍函数，如向上取舍到给定基数的倍数函数“CEILING(数值,基数)”、向下取舍到给定基数的倍数函数“FLOOR(数值，基数)”、向上取舍到最接近的偶数函数“EVEN(数值)”、向上取舍到最接近的奇数函数“ODD(数值)”、四舍五入到基数的倍数函数“MROUND(数值，基数)”（安装数据分析工具后可用）、按小数位数向上取舍函数“ROUNDUP(数值，小数位数)”、按小数位数向下取舍函数“ROUNDDOWN(数值，小数位数)”。注意：CEILING 与 FLOOR 的符号必须相同；ROUND、ROUNDUP、ROUNDDOWN 的小数位数可正可负，如果小数位数为正值，则表示取舍到小数点以右的“小数位数”；如果小数位数为 0，则表示取整数；如果小数位数为负值，则表示取舍到小数点以左的“小数位数”。

7. 求绝对数

无论数据是正还是负，均取正值，数学上表示为$|x|$，如$|-10|=|10|=10$。绝对数函数为：ABS(数值)，如“=ABS(–10)”等于 10。

8. 余数、商

计算被除数除以除数所得余数的函数是：MOD(被除数,除数)，如“=MOD(11,3)”得余

数为 2。

计算被除数除以除数所得商的函数是：QUOTIENT(被除数,除数)，如“= QUOTIENT (11,3)”得商为 3。

安装了数据分析工具后，可用“GCD(数值, [数值,[数值]])”求最大公约数，用“LCM(数值, [数值,[数值]])”求最小公倍数。

9. 连乘与连加

（1）连乘积

如果变量 x 包括 a、b、c、…，则用数学符号表示为 $\Pi x = a \times b \times c \times \cdots$。Excel 计算连乘积的函数为：PRODUCT(a,b,c,…)，如“=PRODUCT(4,3,2)”等于 24。注意：括号中也可以是单元格引用。

（2）累计求和

如果变量 x 包括 a、b、c、…，则用数学符号表示为 $\Sigma x = a + b + c + \cdots$。Excel 计算累计求和的函数为：SUM(a,b,c,…)，如“=SUM(4,3,2)”等于 9。注意：括号中也可以是单元格引用。

Excel 也有一个求平方和的函数：SUMSQ(a,b,c,…)，用数学符号可表示为 $\Sigma x^2 = a^2 + b^2 + c^2 + \cdots$，如“=SUMSQ(4,3,2)”等于 29。注意：括号中也可以是单元格引用。

（3）条件求和

在实际工作中，经常需要根据指定条件求和。Excel 的条件求和函数的形式为：“SUMIF(用于条件判断的单元格区域,条件,求和的单元格区域)”。

该函数包括三个部分：第一部分用于条件判断，第二部分为条件，第三部分为要求求和的区域。其含义为：在第一部分的单元格区域中，如果满足第二部分的条件，那么就对第三部分相应单元格区域求和。例如，已知某医院第一季度的内、外、妇、儿科住院病人数（见图 2-9），现想知道第一季度外科共有多少住院病人。此时，用于条件判断的单元格区域为 B2:B13（科别）；其条件为外科，表格中相应单元格有：B3、B7、B11；要求求和的单元格区域为 C2:C13（住院人数）。因此，第一季度外科共有住院人数可利用“=SUMIF(B2:B13,B3,C2:C13)”进行计算，得结果为 160（见光盘中“02-04 条件求和.xlsx”文件）。



	A	B	C	D	E	F
1	月份	科别	住院人数			
2	1	内科	50	160		
3	1	外科	30			
4	1	妇科	45			
5	1	儿科	53			
6	2	内科	67			
7	2	外科	43			
8	2	妇科	87			
9	2	儿科	32			
10	3	内科	65			
11	3	外科	87			
12	3	妇科	87			
13	3	儿科	91			

图 2-9　条件求和

2.9.2　数组运算

将某个单元格区域内的所有数据看作一个整体，就是数组（array）或矩阵（matrix）。数组运算公式与一般数据运算公式类似，可以直接进行各种数学运算，也可以使用各种函数。其差别主要在以下几个方面：

①计算的对象和结果是一组数据。一般数学运算的对象可以是一组数据，但计算结果只有一个值，而数组的计算结果往往是一组数据；

②按 Ctrl + Shift + Enter 组合键结束输入。一般计算公式编辑完成后，都是按回车（Enter）键或单击编辑栏左侧的“√”结束公式计算。而数组运算公式输入完成后，必须按 Ctrl + Shift + Enter 组合键，这时相应的公式两端会自动加上花括号“{}”，相应的计算也将按数组运算规则进行；

③结果区域是一个整体。建立公式前，应先根据数组判断结果为几行几列，然后在工作表的适当位置，选定相同大小的单元格区域建立数组公式。当按 Ctrl + Shift + Enter 组合键结束输入的数组公式时，计算结果将会填充到选定的单元格区域，这时的结果区域是一个整体，不能单独编辑其中某一个单元格。

（1）数组与数值的计算

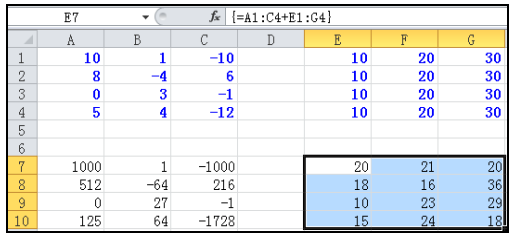
对单元格区域中的每个数据可一次性直接进行加、减、乘、除、乘方、开方等运算。例如，在 A1:C4 单元格区域有 3 列 4 行数据（见图 2-10），对这 12 个数据中的每一个进行 3 次乘幂（立方），可按下列步骤进行：

① 按住鼠标左键拖放选定存放结果的单元格区域，本例选取 A7:C10；

② 在编辑栏输入所需的计算公式“=A1:C4^3”；

③ 按 Ctrl + Shift + Enter 组合键确认（注意不能单独按 Enter 键）。

这时公式编辑栏的公式自动变为“{=A1:C4^3}”，并在单元格区域 A7:C10 获得相应的计算结果（见图 2-10）。



	A	B	C	D	E	F	G
1	10	1	-10		10	20	30
2	8	-4	6		10	20	30
3	0	3	-1		10	20	30
4	5	4	-12		10	20	30
5							
6							
7	1000	1	-1000		20	21	20
8	512	-64	216		18	16	36
9	0	27	-1		10	23	29
10	125	64	-1728		15	24	18

图 2-10　数组的运算

（2）两个数组间的计算

对于两个具有相同行、相同列的矩形区域，可进行相应两数的加、减、乘、除运算。

例如，将图 2-10 中单元格区域 A1:C4 数组与区域 E1:G4 数组相加，可按下列步骤进行：

①按住鼠标左键拖放选定存放结果的单元格区域，本例选 E7:G10；

②输入所需的计算公式“=A1:C4+E1:G4”，按 Ctrl + Shift + Enter 组合键确认。这时公式编辑栏的公式自动变为“{=A1:C4+E1:G4}”，并在单元格区域 E7:G10 获得相应的计算结果（见图 2-10，也见光盘中“02-05 数组运算.xlsx”文件）。

2.9.3 矩阵的运算与求解线性方程组

当进行多元回归分析、求解线性方程组的时候，需要运用矩阵的加、减、乘、求逆、求转置等运算。矩阵运算需要以数组为工具，但它有其自身的运算规则，部分运算需要使用有关矩阵函数来实现。

（1）矩阵的加减法

当两个矩阵相加减时，要求两个矩阵的行数相等、列数也相等，但行数与列数之间可以不等。矩阵的加减法与数组的加减法运算完全相同（见图 2-10）。

（2）矩阵的乘法运算

一个数值与一个矩阵相乘，可按上述数组计算方法运算。

但当两个矩阵间相乘时，要求前一个矩阵的列数等于后一个矩阵的行数；乘积获得的矩阵的行数与前一个矩阵行数相同，列数与后一个矩阵的列数相同，且第 i 行第 j 列元素等于前一个矩阵的第 i 行与后一个矩阵的第 j 列各对应元素乘积之和。例如，前一个矩阵位于单元格区域 A1:C2，为 2 行 3 列；后一个矩阵位于单元格区域 E1:F3，为 3 行 2 列，如图 2-10 所示（也见光盘中“02-06 两个矩阵间相乘.xlsx”文件）。这两个矩阵相乘将获得一个 2 行 2 列的矩阵，如果将乘积获得的矩阵结果放在单元格区域 A5:B6，则 $A5=1\times4+0\times(-1)+3\times2=10$，$B5=1\times1+0\times1+3\times0=1$，$A6=2\times4+1\times(-1)+0\times2=7$，$B6=2\times1+1\times1+0\times0=3$。Excel 具体计算步骤如下：

①选 A5:B6 作为存放结果的单元格区域；

②输入所需的计算公式“=MMULT(A1:C2,E1:F3)”，按 Ctrl + Shift + Enter 组合键确认。这时公式编辑栏的公式自动变为“{=MMULT(A1:C2,E1:F3)}”，并在单元格区域 A5:B6 获得相应的计算结果。

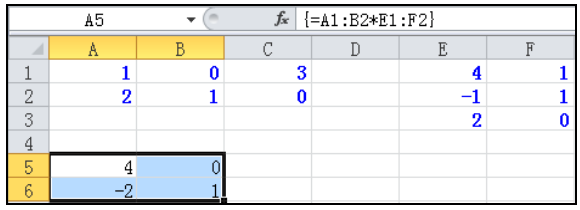


图 2-11 两个矩阵间相乘

温馨提示：两个数组相乘与两个矩阵相乘获得的结果是完全不同的。两个数组相乘要求有相同的行数和相同的列数，它们之间相乘是对应格子的元素相乘。而两个矩阵相乘要求前一个矩阵的列数等于后一个矩阵的行数，其结果的第 i 行第 j 列元素等于前一个矩阵的第 i 行与后一个矩阵的第 j 列各对应元素乘积之和。

（3）逆矩阵

设 $\boldsymbol{A}$ 为 n 阶方阵，$\boldsymbol{E}$ 是 n 阶单位矩阵，若存在 n 阶方阵 $\boldsymbol{B}$，使得 $\boldsymbol{A}\cdot\boldsymbol{B}=\boldsymbol{B}\cdot\boldsymbol{A}=\boldsymbol{E}$，则称 $\boldsymbol{A}$ 为一个非奇异矩阵，并称 $\boldsymbol{B}$ 是 $\boldsymbol{A}$ 的逆矩阵，通常将 $\boldsymbol{B}$ 记为 $\boldsymbol{A}^{-1}$。如果不存在符合以上条件的 $\boldsymbol{B}$（即 $\boldsymbol{A}^{-1}$），则称 $\boldsymbol{A}$ 为奇异矩阵。

例如，$\boldsymbol{A}$ 矩阵为 3 行 3 列，即 3 阶方阵，位于单元格区域 A1:C3（见图 2-12），求该矩阵的逆矩阵 $\boldsymbol{A}^{-1}$。

Excel 具体实施步骤如下：

①按住鼠标左键拖放选定存放结果的单元格区域，本例选 E1:G3；

②在编辑栏输入所需的计算公式“=MINVERSE(A1:C3)”，按 Ctrl + Shift + Enter 组合键确认。这时公式编辑栏的公式自动变为“{=MINVERSE(A1:C3)}”（或=MINVERSE({2,1,1;1,0,2;3,1,2})，其中花括号内为矩阵值，每行值以分号分隔，每行各值以逗号分隔，并在单元格区域 E1:G3 获得相应的计算结果，即逆矩阵 $\boldsymbol{A}^{-1}$（见图 2-12，也见光盘中“02-07 逆矩阵.xlsx”文件）。

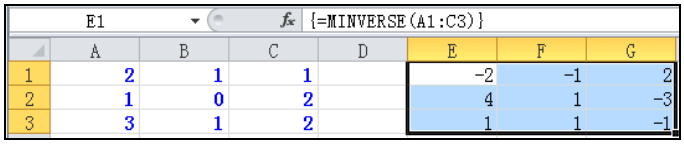


图 2-12　逆矩阵

如果将图 2-12 中的原矩阵 $\boldsymbol{A}$（左侧）与逆矩阵 $\boldsymbol{A}^{-1}$（右侧）相乘，即用计算公式“=MMULT(A1:C3,E1:G3)”可得到单位矩阵

$$\boldsymbol{E}=\begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix}$$

（4）行列式

行列式有如下特点：

①行与列互换，行列式的值不变；

②两行（列）对调，行列式的值改变符号；

③以常数 k 乘行列式的某一行等于以 k 乘这个行列式；

④某行的各元素如果是二数的和，则可将此行列式转化为行列式的和；

⑤某一列的各元素加上另一列的各元素的 λ 倍，行列式的值不变。

依照这些特点可将行列式简化为二阶方阵，求解行列式的值。

Excel 有一个求解行列式值的函数，即 MDETERM 函数，可简单计算行列式的值。如图 2-13 所示单元格区域 A1:D4 的数据，采用公式“= MDETERM(A1:D4)”获得行列式的值为–126（见光盘中“02–08 行列式.xlsx”文件）。

注意：因为所得结果为一个值，而不是一个矩阵，所以可按 Enter 键确认，而不是按“Ctrl + Shift + Enter”组合键确认。

对于图 2-13 所列行列式，也可直接采用将数据代入的方式进行计算，其 Excel 公式为“=MDETERM({1,-1,1,-3;1,2,1,3;2,1,-1,-3;3,-3,-3,-1})”，其中花括号内为行列式值，每行值以分号分隔，每行各值以逗号分隔。

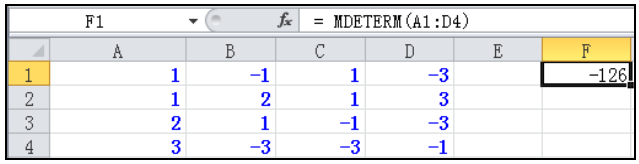


图 2-13　行列式的计算

（5）转置矩阵

如果将矩阵 $\boldsymbol{A}$ 的行列互换，所得矩阵即为 $\boldsymbol{A}$ 的转置矩阵，记为 $\boldsymbol{A}'$。可采用 Excel 的 TRANSPOSE 函数，并按 Ctrl + Shift + Enter 组合键确认。

也可简单地复制该矩阵，然后用选择性粘贴，选取右下角“转置”，并单击“确定”按钮获得。

（6）求解线性方程组

求解下列三元一次方程组：

$$\begin{cases} 5x+6y+7z=8 \\ 6x+15z=6 \\ 2x+6y=7 \end{cases}$$

对于这个方程组可用矩阵表示为：

$$\begin{bmatrix} 5 & 6 & 7 \\ 6 & 0 & 15 \\ 2 & 6 & 0 \end{bmatrix} \times \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 8 \\ 6 \\ 7 \end{bmatrix} \text{或} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 5 & 6 & 7 \\ 6 & 0 & 15 \\ 2 & 6 & 0 \end{bmatrix}^{-1} \times \begin{bmatrix} 8 \\ 6 \\ 7 \end{bmatrix}$$

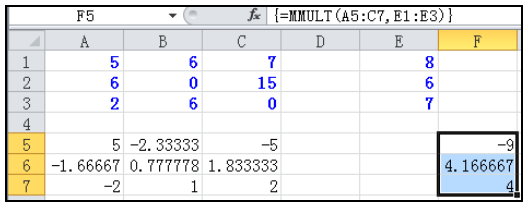
由此可见，要求解该方程组，应先计算系数矩阵的逆矩阵，再将其乘以结果矩阵。

Excel 操作如下：

① 将系数矩阵放在单元格区域 A1:C3，结果矩阵放在单元格区域 E1:E3（见图 2-14，也见光盘中“02-09 求解线性方程组.xlsx”文件）；

② 选定单元格区域 A5:C7，以便存放逆矩阵；在编辑栏输入公式“=MINVERSE(A1:C3)”，按 Ctrl + Shift + Enter 组合键确认。

③ 求得逆矩阵后，再选定单元格区域 F5:F7，用于存放求解方程组的结果；在编辑栏输入公式“=MMULT(A5:C7,E1:E3)”，按“Ctrl + Shift + Enter”组合键确认。此时获得线性方程组的解为 $x=-9$，$y=4.17$，$z=4$。



	A	B	C	D	E	F
1	5	6	7		8	
2	6	0	15		6	
3	2	6	0		7	
4						
5	5	-2.33333	-5			-9
6	-1.66667	0.777778	1.833333			4.166667
7	-2	1	2			4

图 2-14　求解线性方程组

为了验证求解结果，选定单元格区域 G1:G3 存放结果，在编辑栏输入公式“=MMULT(A1:C3,F5:F7)”，按 Ctrl + Shift + Enter 组合键后便获得所需结果。比较单元格区域 E1:E3 与 G1:G3，可见计算结果与结果矩阵完全一致。

2.10　创建自定义函数

在 Excel 中，用户可以创建自定义函数。

比如打算建立一个自定义函数，命名为“cjdjhf”（成绩等级划分），该函数的作用是：成绩 60 分以下时显示为不及格，60 分及以上时显示为及格。创建此自定义函数的具体步骤如下：

① 单击“开发工具→代码→Visual Basic 编辑器”；

② 在弹出的“Visual Basic 窗口”中单击“插入→模块”；

③ 在属性窗口单击 图标，在左下方框中“（名称）”后键入“成绩等级划分”代替“模块 1”；

④ 在出现的空白程序窗口中键入自定义函数 VBA 程序（见图 2-15）。

⑤ 单击“文件→关闭并返回到 Microsoft Excel”，完成自定义函数的创建，保存后即可使用。如在 B1:B9 中键入考试成绩 30、30、…、90，在单元格 C1 中键入“=cjdjhf(B1)”，将此公式拖放至 C9，得到结果为不及格、不及格、…、及格。见光盘文件“cjdjhf.xlsm”。

Excel 将自动检查其逻辑的正确性。此后，可以像使用 Excel 内部函数一样，在这个工作簿内的不同工作表中使用自定义函数 cjdjhf。

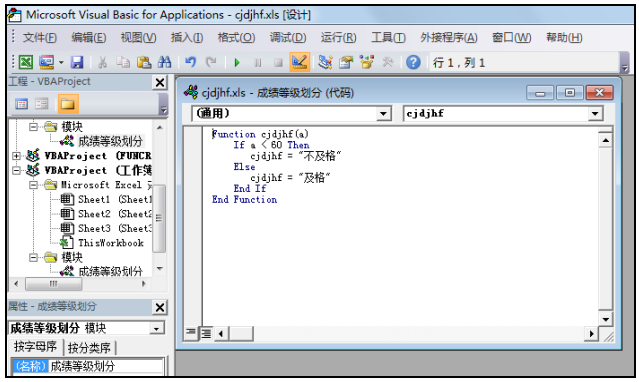


图 2-15　自定义函数

自定义函数需要具备编辑 VBA 程序的知识，对于一般用户难度较大，这里只作简要介绍。

2.11　获取函数帮助信息

当正确输入“=”、函数名以及左括号后，Excel 2003 以上版本会自动提示函数所需的参数，双击函数名还会自动跳出该函数的帮助信息界面。为了获取更详细的信息，还可以单击插入函数“ *fx* ”按钮，弹出“插入函数”对话框（见图 2-16）。在此界面提供了“搜索函数”、“选择类别”、“选择函数”、“有关该函数的帮助”等信息。

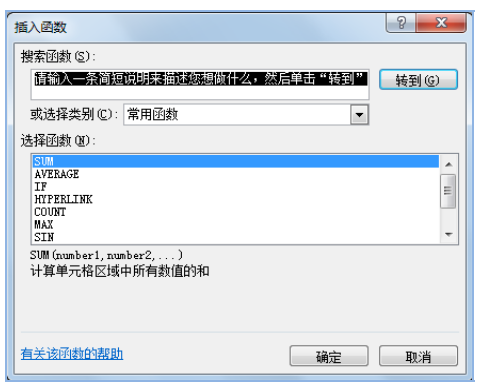


图 2-16　“插入函数”对话框

此外，也可以参考附录 A 以及光盘文件“按类别列出的 Excel 函数.xlsx”与“按字母排序的 Excel 函数.xlsx”的内容。附录 A 中比较详细地列举了与统计分析有关的 7 类函数，依次是：数学和三角、统计、日期与时间、文本、逻辑、查询和引用、数据库。对于统计应用不多的财务、信息、工程、多维数据集函数，对其用途也作了简单介绍。

第 3 章

数据清单的建立与利用

这里所指的“数据清单”，与常用 dBase、FoxBase、Access 数据库管理系统的“数据库”相对应，只是没有数据库的定义那么严谨，适用范围更广泛。

除直接输入数据或调用数据文件的一般功能外，Excel 还可由公式或一般数据快速填充方式产生新的数据，数据的复制、移动、插入、删除、排序、筛选等编辑操作非常灵活，经格式化后的数据工作表美观好看，使得 Excel 成为目前建立中小型数据文件的最常用、最优秀软件之一。由 Excel 产生的数据文件得到了各界的公认，如世界知名统计软件 SAS、SPSS、S-PLUS 等，可以方便地利用 Excel 文件进行统计分析。本章将介绍如何录入、导入数据来建立 Excel 数据文件；为了充分利用 Excel 数据，也将介绍从 Excel 中导出数据，或由其他统计分析软件调用 Excel 数据的方法。

3.1 数字的单元格格式

在输入数据前，如果事先了解并设置好单元格数据格式，输入的数据不仅可按要求显示，而且可限制输入数据时出现的一些错误。

Excel 设置数据格式的方法很简单，只要选中需要设置格式的单元格，单击鼠标右键，在弹出的快捷菜单中选择“设置单元格格式”命令（或单击“开始”下“（单元格）格式”，选取“设置单元格格式”），会弹出“设置单元格格式”对话框（见图 3-1），选中“分类”框下所需要的格式，并在其右边选择或修改必要的选项，单击“确定”即可对单元格数据格式进行修改。

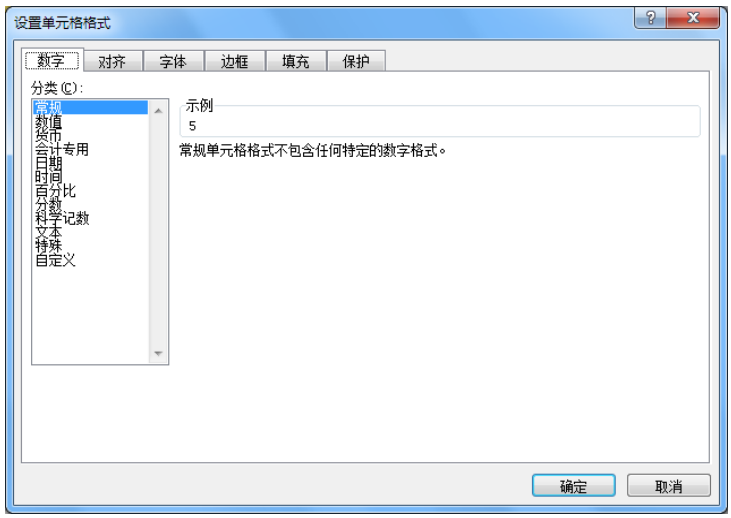


图 3-1　单元格数字格式

可设置的单元格数字格式如下：

（1）常规（general）格式

常规格式是 Excel 默认的单元格格式，新建 Excel 工作表的单元格在格式化前均为常规格式。

对输入的数字而言，只要单元格足够宽，输入和显示的数字就完全相同，并在计算时按数字处理。如果单元格的宽度不足以容纳整个数字，则根据单元格宽度对数字取整，如数字“12345678.99”在单元格宽度为 100 个像素时显示为“12345678.99”，当单元格宽度为 50 个像素时使用科学记数法，显示为“1E+07”。如果单元格宽度继续减小（如 40 个像素），则显示“####”，此时只要增大单元格宽度就可以正确显示数字。

（2）数值（number）格式

可对“小数位数”、“是否使用千位分隔符”和“负数”的显示格式进行设置。如果某一数字小数点右侧的位数大于所设置的小数位数，该数字将按您的设置进行四舍五入。一般情况下的负数前为负号（-），但如果需要，也可将负数以红色（赤字）、红色带括号、黑色带括号等多种形式显示。

（3）货币（currency）格式

它可由用户对“小数位数”、是否添加货币符号和负数的显示格式进行设置。如果要让 Excel 使用默认的货币符号，可在启动 Excel 之前，打开“控制面板→时钟、语言和区域→更改日期、时间或数字格式”，在弹出的“区域和语言”对话框中“格式”选项卡下选择“其他设置”，在弹出的“自定义格式”对话框中，选择要使用的货币符号、显示位置等。

（4）会计专用（accounting）格式

它与货币格式基本相同，只是不能对负数显示格式进行设置。

（5）日期（date）格式

如果是在一个“常规”单元格内输入日期，则输入与显示的结果完全相同。例如输入“2007-12-6”，则显示“2007-12-6”。若事先或事后设置了单元格的日期格式，则输入内容会按设置的格式转换为设定格式，如“二〇〇七年十二月六日”、“2007 年 12 月 6 日”等选定的格式。若输入了不合理的日期，如“2007-2-30”，则不会进行转换，Excel 只将该输入值当作文本（单元格左对齐），而不当作日期（单元格右对齐）。

如果将包含数字（称为序列数）的单元格设置为日期，则 Excel 会以 1899 年 12 月 31 日为起点，即数字 0 对应 1899 年 12 月 31 日（计算机显示“1900-1-0”），1 对应 1900 年 1 月 1 日，365 对应 1900 年 12 月 30 日，36525 对应 1999 年 12 月 31 日。

（6）时间（time）格式

如果在“常规”单元格内输入时间，则输入与显示的结果也会完全相同。例如输入“15:38”，则显示“15:38”。如果事先或事后为单元格设置了时间格式，则输入“15:38”后可按设置的格式显示为“3:38pm”或“15 时 38 分”等形式。

（7）百分比（percentage）格式

如果输入前单元格为“常规”格式，若在输入的数字后加上%，则自动将所在单元格设置为百分比格式。如果要将任何一个数字格式的单元格设置为百分比格式，则 Excel 会将单元格中的数值乘以 100，并以百分比形式显示。例如“5”的百分比格式为“500.00%”。可以像设置一般数字一样，为百分比格式设置小数位数。

（8）分数（fraction）格式

Excel 分数的分母格式有一位数、二位数、三位数、2、4、8、16、10、百分之几的分数格式选项。例如，选分母为 10 的分数格式，则在设置了该格式的单元格中键入 0.8 显示

为 8/10，键入 1.8 显示“1 8/10”，键入 1.85 显示为“1 9/10”，键入 100 显示“100”。



注意：键入分数时，前面应键入 0 并加上一个空格，然后输入数据，否则 Excel 会误以为是日期数据或文本数据。

（9）科学记数（scientific）格式

如果输入前单元格为“常规”格式，可以直接以科学记数格式进行输入。例如，要输入“10^{18}”可以直接键入“1e＋18”或“1.00E＋18”等，正指数时 E 后面是加号，负指数时为减号，加减号后的数字决定指数的位数。如果事先或事后将单元格设置为科学记数格式，则 Excel 可以自动进行转换。例如，在设有科学记数格式的空白单元格中输入“98000000”，按 Enter 键后自动显示为“9.80E＋07”，其中 E 前面数字的小数位数可事先确定。

（10）文本（text）格式

在“常规”单元格中，如果输入文字或文字与数字混合，则 Excel 自动将其作为文本显示（左对齐），且不参与 SUM 函数等公式的运算。如果输入的是数字，且事后将单元格设置为“文本”格式，Excel 仅将其作为文本显示（左对齐），但在进行加减、自动求和及用 SUM 公式计算时，Excel 仍将它视为数字。如果输入前将单元格设置为“文本”格式，那么 Excel 不仅将其作为文本显示（左对齐），而且不能参与加减、自动求和及用 SUM 公式计算。故输入“病案号”、“科别”等内容时应采用后一种方法。即事先将空单元格设置为文本格式，然后在带文本格式的单元格中输入数字。

如果在要输入的数字前加上一个英文格式的单引号“’”，则输入的数字就是文本格式的数字。

（11）自定义（custom）格式

①自定义数字格式

在“设置单元格格式”对话框中的“数字”选项卡下，选中左边分类框下的“自定义”项，右边的“类型”下显示的均为自定义数字格式。其中包括“0”格式、“＃，＃＃0．00；[红色]-＃，＃＃0．00”格式等，前者可将输入的小数自动取整，后者可在输入千位数时自动加上分隔符且保留两位小数，若输入千位负数则自动加上负号和分隔符并保留两位小数。用户可以以上述格式为基础，创建适合自己特殊用途的数据格式。

②简化文字输入

“@”格式可简化文字输入。格式有“文字@”、“@文字”和“文字@文字”3 种。如果选中“自定义”的“@”格式，在“类型”下的编辑框内“@”之前键入“第”，“@”之后键入“人民医院”，完成“第@人民医院”输入后单击“确定”按钮，以后只要简单输入替换“@”的文字（或数字），便可得到一串文字。如输入“五”，回车后就会显示为“第五人民医院”。如果不再需要上述格式，单击单元格格式的“数字”选项卡中“自定义”（左）

下方的“删除”按钮，即可删除该单元格自定义格式。

③条件格式

注意，Excel 2003 的条件格式在“格式”菜单下寻找；在 Excel 2010 中可以直接在“开始”选项卡下“样式”命令组中找到“条件格式”。在“突出显示单元格规则”选项下，Excel 原始自带的有 8 个条件，即介于（between）、未介于（not between）、等于（equal to）、不等于（not equal to）、大于（greater than）、小于（less than）、大于等于（greater than or equal to）、小于等于（less than or equal to）来定义“单元格数值”的格式；还可单击“其他规则”来新建“条件格式”。

例如，选中需要设置格式的单元格，单击“开始→样式→条件格式▼→突出显示单元格规则→其他规则”，弹出“新建格式规则”对话框（见图 3-2）。将“编辑规则说明”下第一个框选择为“单元格值”，第二个框修改为“未介于”，在后面两个框内分别输入 50 和 50000 后，单击“格式”按钮，在对话框中选择条件成立时数据显示的格式，如字体颜色为红色、字体样式为粗体等。这时，所有住院费用满足“<50”（不包括 50）或“>50000”（不包括 50000）的患者会以红色黑体显示。

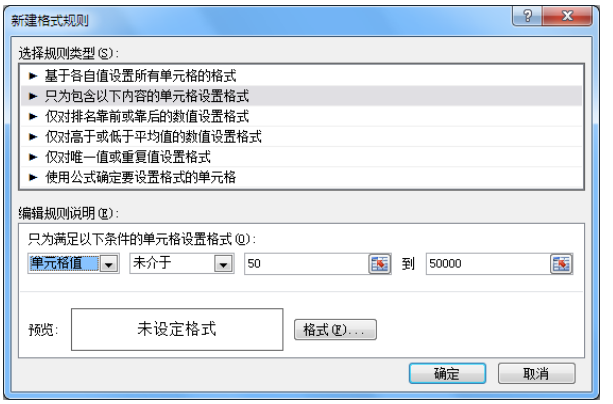


图 3-2 “条件格式”对话框

为了实现数据单元格隔行显示不同的颜色，可采用公式“=AND(LEN($A1)>0, MOD(ROW(),2)=1)”。如果先选择 A1 单元格，按图 3-2 设置条件为真时的颜色，然后修改左侧“单元格数值”为公式，其右侧空白处键入公式“=AND(LEN($A1)>0, MOD(ROW(),2)=1)”，单击“确定”按钮即可。公式含义是：如果当前单元格行数（ROW()）与 2 之比的余数等于 1（MOD(ROW(),2)=1），且当前单元格“$A1”的字符个数大于 0（LEN($A1)>0），两项条件均满足，则采用事先设定的“条件格式”颜色。

（12）特殊（special）格式

在其下面有“区域设置（国家/地区）”选项，可选择不同的文字（如中文）与所在的

区域（如中国）。“中文（中国）”有 3 种特殊格式类型（邮政编码、中文小写数字、中文大写数字，其显示方式分别为：510080、五十一万〇八十、伍拾壹万零捌拾）。选择不同的“区域设置（国家/地区）”可能有不同的特殊格式，例如选择英文（美国），则出现“Zip Code、Zip Code +4、Phone Number、Social Security Number”4 种特殊格式。

3.2 数据的有效性设置

在录入数据前，为了保证数据的有效性（data validation），可对有关单元格进行数据有效范围的设置。在录入数据时，Excel 会对设置了数据有效性范围的单元格进行自动检查，如果录入数据超出了设置的有效范围，Excel 会弹出出错信息提示用户。如性别只有“男”、“女”；ABO 血型只有 A、B、AB、O 型；人的年龄一般不超过 120 岁，且肯定为正数；人的体温一般不超过 45℃。如果超出了规定的范围，应属于无效数据。

1. 设置数据的有效性范围

按下列步骤设置数据的有效性。

① 选中所需设置的单元格；

② 单击“数据→数据工具→数据有效性”，将弹出一个“数据有效性”对话框；

③ 在“设置”选项卡中，将鼠标移至“任何值”右侧的倒三角按钮处，单击左键出现如图 3-3 所示的界面；

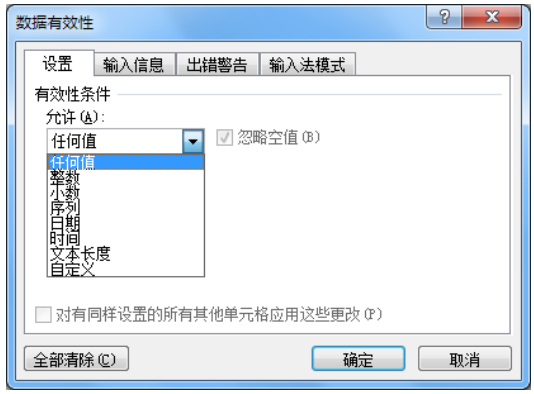


图 3-3　数据的有效性设置对话框

④ 根据数据类型适当选取允许列表框中的有效数据类型，设置有效数据格式。

有效数据类型有：任何值（any value）、整数（whole number）、小数（decimal）、序列（list）、日期（date）、时间（time）、文本长度（text length）、自定义（custom）。“自定义”允许使用公式、表达式或引用其他单元格中的数值来判定输入数值的正确性，其公式得出

的结果必须是“True”或“False”。如果类型是“任何值”，那么对数据将不做任何限制。下面举例说明设置方法。

（1）创建下拉菜单

如果数据值可以分为多类，则可以选择“序列”创建下拉菜单，定义单元格数据的有效性。例如，为了输入 12 个月的“月份”，可按下列步骤操作。

①选择某一单元格区域存放可能的序列分类值，如在单元格区域 A1:A12 存放“一月”、…、“十二月”；

②选中需要设置的单元格区域，如整个第 D 列（即 D:D）；

③单击“数据→数据工具→数据有效性”，将弹出一个“数据有效性”对话框；

④在“设置”选项卡下，将鼠标移至“任何值”右侧的倒三角按钮处单击左键后，再单击“序列”；

⑤单击“来源”框右侧，选取存放序列分类值的区域，本例为 A1:A12（或直接在“来源”框输入“$A$1:$A$12”）（见图 3-4），单击“确定”；

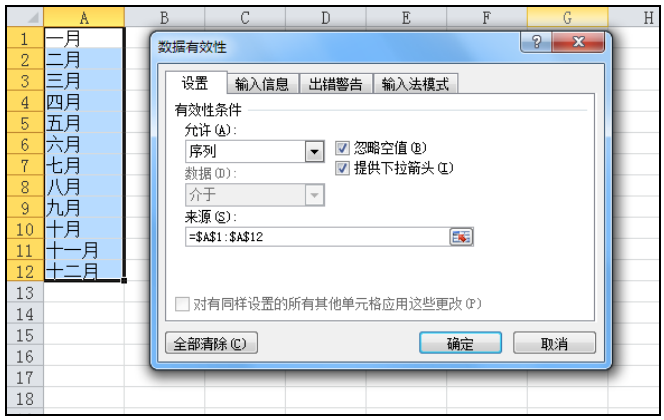


图 3-4　创建下拉菜单

⑥在设置好的单元格区域中单击某一单元格，会在该单元格右侧出现一个倒三角按钮，单击倒三角按钮，会出现事先输入的序列分类值（一月、…、十二月），用鼠标选择其中之一即可。其他单元格也按此操作，可快速完成输入这类数据（见光盘文件“03-01 数据有效性下拉菜单.xlsx”）。

（2）有效数据范围的设置

为了调查广州某社区老年人的生存质量，选择该社区 100 名 60 岁以上老人作为调查对象。欲将年龄数据放在单元格区域 B2:B100。为了设置单元格中年龄的有效值范围，可先选择单元格区域 B2:B100，单击“数据→数据工具→数据有效性”，在弹出的对话框中选择“设置→允许→整数”，在最小值栏输入 60，在“最大值”栏输入 120，单击“确定”即可

完成设置。在设置的单元格中只能输入 60～120 之间的数据（见光盘文件“03-02 数据有效性区间数据的设置.xlsx”），如果输入的数据小于 60 或大于 150，如输入 59，则屏幕会弹出一个对话框，提示“输入值非法”（见图 3-5）。这样可避免输入数据出现的一些错误。

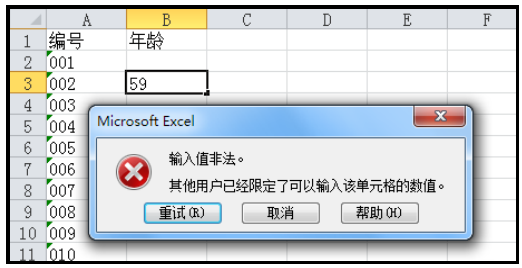


图 3-5　输入值出错的提示信息

（3）由公式定义数据的有效性

由公式定义数据的有效性尽管复杂，但更实用。这里的公式的最终返回结果应该是“TRUE”或“FALSE”。比如，假定 A1 为活动单元格，可由公式“=ISTEXT(A1)”迫使所选单元格区域 A1:A20 只能输入文本，而不能输入数字；假定 B1 为活动单元格，由公式“=COUNTIF($B$1:$B$20,B1)=1”迫使所选单元格区域 B1:B20 只能输入不重复的数字，如果输入的数字有重复，则会弹出错误提示。

实现只能输入文本的有效性设置步骤如下：

①选择需要设置有效性的单元格区域，如本例选 A1:A20；

②单击单击“数据→数据工具→数据有效性”，在弹出的对话框中选择“设置→（允许）自定义”；

③在公式框中写入“=ISTEXT(A1)”（意思是：单元格 A1 如果是文本，则返回“TRUE”，否则返回“FALSE”。因为公式中当前单元格 A1 是相对引用，所以当活动单元格为 A2 时，公式会自动改变为“=ISTEXT(A2)”，…，当活动单元格为 A20 时，公式会自动改变为“=ISTEXT(A20)”，…），单击“确定”按钮即可。

实现只能输入不重复数据的有效性设置步骤如下：

①选择需要设置有效性的单元格区域，如本例选 B1:D20；

②单击单击“数据→数据工具→数据有效性”，在弹出的对话框中选择“设置→（允许）自定义”；

③在公式框中写入“=COUNTIF($B$1:$D$20,B1)=1”，单击“确定”即可（见光盘文件“03-03 数据有效性公式自定义.xlsx”）。

2. 删除单元格的有效性范围设置

如果用户不需要某些单元格设置的有效数据范围，可以将其删除。删除的方法是：选

中需要删除的单元格区域，单击“数据→数据工具→数据有效性”，在弹出的对话框中选择“设置→全部清除”，然后单击“确定”即可。

3. 圈释无效数据

如果数据已经存在，且对某区域设定了数据“有效性”，则可采用“数据→数据工具→数据有效性▼→圈释无效数据”，可对该区域的无效数据给予红线圈释。如果在“数据有效性▼”下选择“删除无效数据标识圈”，则可删除红线圈。

4. 语音朗读

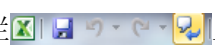
在 Excel 2003 中可以采用“视图→工具栏→文本到语音”，Excel 会弹出

，从左到右的按钮分别是“朗读单元格”、“停止朗读”、“按行”、“按列”、“按回车键朗读”。通过朗读可以帮助发现无效数据。

然而在 Excel 2007 及 2010 版本中语音朗读的功能默认标签在功能区中都找不到，需要用户自定义设置，将其加到快速访问栏。

① 单击“文件→选项界面→快速访问工具栏”，在“从下列位置选择命令”下的第一个框中选择“不在功能区的命令”。

② 在筛选出的命令中单击“按 Enter 开始朗读单元格”，左键单击“添加”将命令添加到快速访问栏中（见图 3-6），按“确定”保存设定并退出选项。

- 回到工作簿窗口，单击快速访问栏上的“按 Enter 开始朗读单元格”即可使用朗读功能。

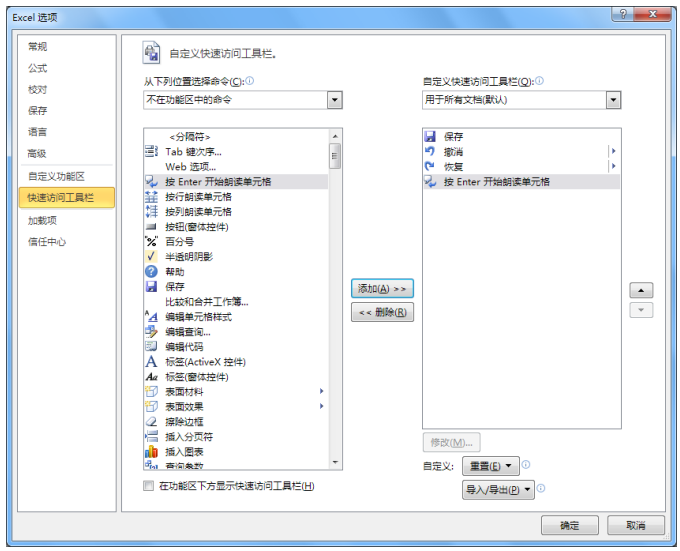
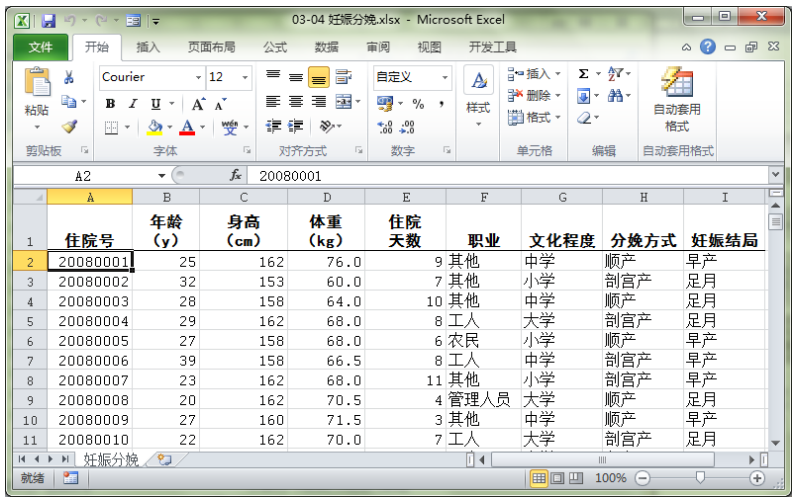


图 3-6　激活朗读选项

3.3 建立数据清单

录入数据时，一般以每一列为一个字段（field），如光盘文件“03-04 妊娠分娩.xlsx”的第 1、2、3、…列分别为住院号、年龄、身高、…，数学或统计学称字段为变量（variable）；以每一行为一个记录（record）（如第 2、3、4 行分别是住院号为 20080001、20080002、20080003、20080004 的记录，数学或统计学称一个记录为一个观测个体（subject 或 individual）或观测单位（observed unit），如图 3-7 所示。



	A	B	C	D	E	F	G	H	I
1	住院号	年龄（y）	身高（cm）	体重（kg）	住院天数	职业	文化程度	分娩方式	妊娠结局
2	20080001	25	162	76.0	9	其他	中学	顺产	早产
3	20080002	32	153	60.0	7	其他	小学	剖宫产	足月
4	20080003	28	158	64.0	10	其他	中学	顺产	足月
5	20080004	29	162	68.0	8	工人	大学	剖宫产	足月
6	20080005	27	158	68.0	6	农民	小学	顺产	早产
7	20080006	39	158	66.5	8	工人	中学	剖宫产	早产
8	20080007	23	162	68.0	11	其他	小学	剖宫产	早产
9	20080008	20	162	70.5	4	管理人员	大学	顺产	足月
10	20080009	27	160	71.5	3	其他	中学	顺产	早产
11	20080010	22	162	70.0	7	工人	大学	剖宫产	足月

图 3-7　数据清单示例

（1）建立数据清单的注意事项

建立数据清单应注意如下几点：

①在一张工作表中，只建立一个数据清单。数据清单中的某些处理命令，如筛选、排序等，每次只能在一个数据清单中使用。

②在数据清单中应避免出现空白整行或空白整列。

③以数据清单的第 1 行创建字段（变量）名。这样做的好处有很多，如执行排序、筛选等命令时，Excel 将自动识别字段名；导出形成其他数据库文件时不易发生错误；其他统计软件调用时方便容易。

④为了将字段名和数据清单数据区分开，可对整个数据清单进行格式化，美化工作表界面。

（2）常规数据录入方式

常规数据录入方式有 3 种：

①单击目标单元格直接输入数据。

②双击目标单元格录入数据。双击目标单元格后，会出现插入光标，鼠标也将变成插入指针“Ⅰ”，将光标移到所需位置录入数据。这种方法多用于修改单元格中的数据。

③单击目标单元格，在编辑栏中输入数据。在 Excel 工作表的编辑栏中，可以复制、粘贴进行编辑或修改数据。在一个单元格中，如果需要录入很多文字，或输入需要编辑的 Excel 公式时，常需采用这种方式。

为了竖向输入数据，每输完一个单元格后应按 Enter 键，此时光标自动向下移动一格；如果打算横向输入数据，每输完一个单元格后应按 Tab 键，此时光标向右移动一格。

大多数情况下，采用 Tab 键、数字小键盘和 Enter 键，可完成数字数据的输入。在需要输入数据的单元格中利用数字小键盘输入数据，按 Tab 键，光标右移一格；继续输入数据，按 Tab 键，光标再右移一格；…；按 Enter 键，光标下移一行，并自动回到最左侧输入数据的单元格位置。继续重复上述操作。这样输入数字方便快捷，且不易出错。

（3）常规数据录入应注意的事项

常规数据录入应注意以下几点。

①最大容量：Excel 2010 一个单元格最多可容纳 32767 个字符。在编辑栏可以显示全部字符，而一个单元格内最多只可以显示 1024 个字符。

②日期的录入：录入年、月、日日期数据时，可按“年-月-日”格式输入，如“2013 年 4 月 1 日”可录入“2013-4-1”；录入月、日时，可按“月-日”格式录入，如“4 月 1 日”可录入“4-1”。也可采用“年/月/日”或“月/日”格式录入。时间可按“小时:分钟”或“小时:分钟:秒”方式，按 24 小时制录入。

③分数的录入：录入数学上的分数数据前，必须在其分数前面加上零和空格，即“0 ”，如输入分数值“3/5”，须键入“0 3/5”，然后按 Enter 键。否则会当作日期处理，应特别给予注意。

④负数的输入：负数可采用两种输入方式，一是输入减号“–”加数字，如“-1”；二是将数字置于括号中，如“(1)”。

⑤公式的输入：输入公式前，必须先按等号“=”，等号意味着输入公式开始（详见第 2 章）。

⑥特殊符号的输入：为了输入特殊符号，可单击“插入→符号→符号”，在选项列表框中选择所需符号，然后单击“插入”。

⑦对齐方式：在默认情况下，单元格中的文本左对齐，而数字、日期、时间等却是右对齐。为了改变对齐方式，可在“设置单元格格式”对话框中，或在功能区“开始”选项卡下的“对齐方式”命令组中进行设置。

⑧让某单元格中的文字或数字转行：如希望“身高（cm）”中的“身高”在单元格上

面，而其度量衡单位“（cm）”在单元格下面，可在输入完“身高”后，按“Alt+Enter”键，光标便转行，然后再输入“（cm）”即可。此点对建立统计表格很有用。

3.4 数据“记录单”的作用

Excel 提供的数据“记录单”具有浏览记录、添加记录、修改记录、删除记录的几个作用。使用“记录单”命令，可以在数据清单中一次输入、显示、查找或删除一行完整记录（观测个体）。

启用“记录单”的操作步骤如下：

① 单击数据清单（见光盘文件“03-04 妊娠分娩.xlsx”）中的任意一个单元格；若数据清单不符合上述“数据清单”的要求，则需采用鼠标选中包含字段（变量）名的单元格区域；

② “记录单”功能在 Excel 2010 默认功能区找不到，需要用户仿照 3.2 的第 4 点添加“语音朗读”功能的步骤，通过 Excel 选项添加到自定义快速访问工具栏中。添加成功后单击自定义快速访问栏中的“记录单”标识符，弹出如图 3-8 所示的对话框。对话框右上方显示了记录单数据在数据清单中的位置，如“1/1402”表示该数据清单共有 1402 个记录，记录单数据为数据清单中的第 1 个记录。

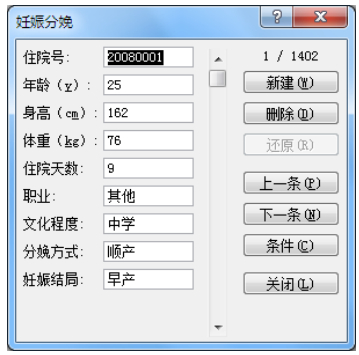


图 3-8 数据清单与“记录单”对话框

在记录单中，可实现如下功能。

（1）浏览数据清单内容

通过单击“上一条”、“下一条”，或拖动对话框中间的垂直滚动条，可快速浏览记录内容。

（2）浏览符合条件的记录

单击对话框中的“条件”，记录框数据将变成空白。此时，在对话框的任意字段框内给

出“条件”，可查找符合条件的记录。“条件”可以在一个框内给出，也可以在多个框内给出。例如，查找身高为 169cm，住院天数为 5 天的记录，可在单击“条件”按钮后，在字段名为“身高”的框内键入 169，在字段名为“住院天数”的框内键入 5，按 Enter 键后，便可找出其中一个（住院号为 20080618）符合条件的记录。单击“上一条”或“下一条”按钮可找到其他符合条件的记录，如本例单击“下一条”按钮，得到住院号为 20080637 的记录。

（3）添加记录

单击“新建”，记录框数据变成空白，且右上角显示“新建记录”字样。此时，用户可将鼠标指向每一个空白位置，逐一输入。每输完一个字段不要使用 Enter 键，而应使用鼠标指向另一个字段名的空白框，并输入新内容，直至您认为输完新增记录的所有字段后，才能按 Enter 键，并进入下一个记录的添加框。单击“上一条”按钮回到浏览状态；单击“关闭”结束记录单的操作。

在添加记录的过程中，若单击“还原”，则将当前录入记录的内容全部清除，仍处于添加当前新记录的状态。

（4）修改记录

在浏览记录内容时，可同时对记录的字段进行修改。修改完毕后，单击“新建”、“关闭”按钮或按 Enter 键，系统将接受修改。单击其他按钮，将放弃所作的修改。

（5）删除记录

在浏览记录内容时，单击“删除”，则删除正显示的当前记录。

在任何时候单击“关闭”按钮，将结束记录单的操作，回到数据清单的工作表界面。

3.5　数据的自动填充

在输入数据或公式的过程中，如果输入的数据或公式具有一定的规律性，可以利用 Excel 的自动填充功能帮助完成数据或公式的输入。这样既减少了数据输入的工作量，又避免了数据错误输入的概率，大大提高了工作效率。这一优点也许是 Excel 备受用户欢迎的重要功能之一。

3.5.1　简单数据的自动填充

简单数据的自动填充可采用“填充柄”、“‘Ctrl+Enter’组合键”或“填充序列对话框”来实现。

1. 填充柄

一般情况下，Excel 工作表中的鼠标显示为空心十字。当用鼠标单击一个单元格或用鼠

标以拖放方式选定一个连续的单元格区域时，选定的单元格或区域的外围框线的右下角会出现一个黑点，这个黑点就是填充柄“ ”。将鼠标移向该点，鼠标由空心十字改变为实心十字。此时按住鼠标左键，拖放至需要填充的单元格，松开鼠标左键，可完成自动填充。

当在某一个单元格输入一个日期、时间、星期、“文本+数字”或“数字+文本”时，使用填充柄可以对选定单元格内容进行自动填充（见图 3-9 中的第 A、B、C、D、E 列，也见光盘文件“03-05 数据自动填充规律.xlsx”）。

	A	B	C	D	E	F	G	H	I	J	K
1	2007/12/10	0:59	星期一	a1	2007年	a	100	100	a	中国	1
2	2007/12/11	1:59	星期二	a2	2008年	a	100	95	b	美国	2
3	2007/12/12	2:59	星期三	a3	2009年	a	100	90	a	日本	4
4	2007/12/13	3:59	星期四	a4	2010年	a	100	85	b	中国	5.333333
5	2007/12/14	4:59	星期五	a5	2011年	a	100	80	a	美国	6.833333
6	2007/12/15	5:59	星期六	a6	2012年	a	100	75	b	日本	8.333333
7	2007/12/16	6:59	星期日	a7	2013年	a	100	70	a	中国	9.833333
8	2007/12/17	7:59	星期一	a8	2014年	a	100	65	b	美国	11.33333
9	2007/12/18	8:59	星期二	a9	2015年	a	100	60	a	日本	12.83333
10	2007/12/19	9:59	星期三	a10	2016年	a	100	55	b	中国	14.33333
11											

图 3-9　规律性变化的数据

利用填充柄填充一个一般的文本或一个数字，只能在每个单元格填充相同（不变）的内容（见图 3-9 中的第 F、G 列）。

当输入两个数值型数据时，选定这两个单元格作为当前单元格区域，拖放单元格区域右下角的填充柄，Excel 会按这两个单元格的数值及其差值自动填入等差序列（见第 H 列）。如果先输入的是两个文本值（如 a、b），则只能重复这两个单元格的文本内容（见第 I 列）。同样，如果先输入三个文本值（如中国、美国、日本，单元格分别为 J1、J2、J3），选取单元格区域 J1:J3，拖放区域右下角的填充柄，则可获得这三个单元格的重复文本内容（见第 J 列）。如果在单元格 K1、K2、K3 中分别输入 1、2、4（等比数列），选取单元格区域 K1:K3，拖放区域右下角的填充柄，Excel 将无法自动识别这一“等比数列”的内容（见第 K 列）。

2. “Ctrl+Enter”组合键

为了在不同单元格中一次性快速输入同一数据（数值、文本、逻辑值等），可按如下步骤操作。

① 选定单元格区域。如果要选择几个不连续的单元格区域，在选取了第一个单元格区域后，再按住 Ctrl 键继续选取其他一个或几个单元格区域；

② 键入数据，如“Excel 2010”；

③ 然后按“Ctrl＋Enter”组合键。

例 3-1：请在单元格区域 A1:A6、B1:F1、C3:D6、F4:F6 中快速输入文本“Excel 2010”。

解：选取单元格区域 A1:A6，按住 Ctrl 键，继续选取 B1:F1、C3:D6、F4:F6，松开 Ctrl

键，采用键盘输入文本“Excel 2010”，在按住 Ctrl 键后再按 Enter 键，即可按要求将文本“Excel 2010”一次性填放入单元格区域 A1:A6、B1:F1、C3:D6、F4:F6 中（见图 3-10）。

	A	B	C	D	E	F
1	Excel 2010	Excel 2010	Excel 2010	Excel 2010	Excel 2010	Excel 2010
2	Excel 2010					
3	Excel 2010		Excel 2010	Excel 2010		
4	Excel 2010		Excel 2010	Excel 2010		Excel 2010
5	Excel 2010		Excel 2010	Excel 2010		Excel 2010
6	Excel 2010		Excel 2010	Excel 2010		Excel 2010
7						

图 3-10　Ctrl＋Enter 组合键的自动填充

3. 填充序列对话框

在一个单元格输入第一个数值，选定一个单元格区域，使输入的第一个数值位于单元格区域的顶行、底行、最左边或最右边，单击编辑菜单中的“填充”，会弹出一个对话框，根据第一个数值所在的位置选择向下填充、向上填充、向右填充、向左填充命令进行填充。

3.5.2　等比数列数据的填充

前面介绍“填充柄”时提到，等差数列数据在给出前两个数值后可采用“填充柄”完成（见图 3-9 中第 H 列），但等比数列数据则不能（见图 3-9 中第 K 列），为了进行等比数列的填充可采用下列步骤操作。

① 输入第一个数值，然后选中需要填充的单元格区域。例如，先在 A1 单元格输入 1，然后单击列标 A，选中第 A 列；

② 单击“开始→编辑→填充▼→系列”，弹出“序列”对话框；

③ 在“序列”对话框中，根据需要指定序列的各项参数，并单击确定按钮。例如，在“序列产生在”下选取“列”，“类型”下选取“等比序列”，步长值定为 2，终止值定为 10000，单击“确定”按钮后，可在单元格区域 A1:A14 产生 1、2、4、8、16、32、64、128、256、512、1024、2048、4096、8192 的等比数列数据（见图 3-11）。

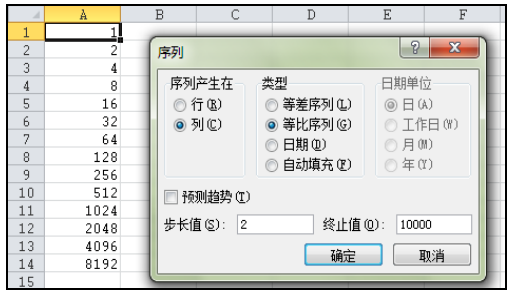


图 3-11　等比数列参数的选取

产生等比数列必须注意两点：

①等比数列的第一个数值不能为 0，因为 0 除任何数为 0；

②只能按行或列产生数据，不能同时按行和列产生数据。

当然，图 3-11 也可用来产生等差数列，读者可仿照上面操作实现。

3.5.3 公式的自动填充

在第 2 章“公式与函数”中，介绍了公式的单元格相对引用、绝对引用与混合引用，提到对公式进行复制或填充。公式自动填充可利用相对引用、绝对引用与混合引用的不同特点，采用“填充柄”，快速建立有规律的单元格引用公式。

公式自动填充的操作步骤如下：

① 选定某单元格，在该单元格中建立公式（参见第 2 章）；

② 选定公式所在单元格，将鼠标移至该单元格的右下角（即填充柄）处，鼠标由空心十字改变为实心十字后，按下鼠标左键，向需要建立公式的单元格方向拖动，到最后一个需要建立公式的单元格，松开鼠标左键。



注意：

①如果是相对引用，拖放的公式将根据参考位置的纵横坐标的改变发生相对的改变。如在 E1 单元格建立公式，相对引用 A1 单元格，如果将单元格 E1 拖放至单元格 F1（向右拖放），则 F1 中引用的 A1 改变为 B1；如果将 E1 拖放至 E2（向下拖放），则 E2 中引用的 A1 改变为 A2。

②如果是绝对引用（列标、行号前均加上$符号），在单元格 E1 建立公式，绝对引用$A$1 单元格，无论将公式向哪个方向移动（比如向上、下、左、右拖放），公式中$A$1 始终保持不变。

③如果是混合引用$A1（列标绝对引用、行号相对引用），将 E1 拖放至 F1（向右拖放），则 F1 中引用的$A1 不改变；如果将 E1 拖放至 E2（向下拖放），则 E2 中引用的$A1 改变为$A2。如果是混合引用 A$1（列标相对引用、行号绝对引用），则向（左）右移动改变，向（上）下移动不改变(详见 2.5 节)。

3.5.4 智能标记的应用

图 3-9 中的第 A、B、C、D、E 列显示了填充时的一般改变规律（见 3.5.1 节“1. 填充柄”），我们是否可以改变这些规律，而按我们的意愿进行自动填充呢？Excel 提供了“自动填充选项”，可以实现按意愿填充。

例如，当在单元格 A1 中输入 2011-1-31 后，将鼠标移至右下角填充柄处，拖放至 A2

后，A2 单元格将自动显示“2011-2-1”（见图 3-12 左），如果想让显示每个月月底的日期，可在“自动填充选项”中选取“以月填充”（见图 3-12 中），然后进行拖放，如拖放至 A14，可以看出 2011 年及 2012 年 1、2 月的每月月底日期。由此可见，2 月只有 28 天，而 2012 年的 2 月有 29 天（见图 3-12 右）。



图 3-12　智能标记填充

3.6　数据的分列

由很多软件产生的结果数据，复制、粘贴到 Excel 工作表后，一列数据不是一个变量的数据值，而是多个变量的数据值。例如，由著名统计软件 SAS 9.13 的“PROC PRINT;”输出的“03-04 妊娠分娩.xlsx”Output 窗口结果数据（见光盘文件“03-04 妊娠分娩.txt”），如果将其中数据粘贴到 Excel 后，每行数据的多个变量值只占一个单元格（尽管显示时占据多个单元格的位置）。

注意：在 Excel 2010 改进了分列功能，如果直接打开文本文件，会自动进入分列功能，很大程度增强了 Excel 调用其他类型护具文件的便捷性。

例 3-2：在图 3-13 中，变量“Obs、_COL0、_COL1、_COL2、_COL3、_COL4、_COL5、_COL6、_COL7、_COL8”均存放在第 A 列，A3 单元格包括“Obs、_COL0、_COL1、_COL2、_COL3、_COL4、_COL5、_COL6、_COL7、_COL8”10 个变量名，A5 单元格包括“1、20080001、25、162.0、76.0、9、其他、中、顺产、早?”10 个数字。请将该数据转变成正常的 Excel 数据。

解：为了将数据正常分列，即一个变量只占据一列，可按下列步骤进行操作。

① 选取需要分列的单元格区域。如本例粘贴“03-04 妊娠分娩.txt”文件被数据后，选取整个 A 列；

A1 The SAS System 15:10 Saturday, December 11, 2007 1″

	A	B	C	D	E	F	G	H	I	J
1	The SAS	System	15:10 Saturday, December 11, 2007			1″				
2										
3	Obs	_COL0	_COL1	_COL2	_COL3	_COL4	_COL5	_COL6	_COL7	_COL8
4										
5	1	20080001	25	162.0	76.0	9	其他	中	顺产	早?
6	2	20080002	32	153.0	60.0	7	其他	小	剖宫	足?
7	3	20080003	28	158.0	64.0	10	其他	中	顺产	足?
8	4	20080004	29	162.0	68.0	8	工人	大	剖宫	足?
9	5	20080005	27	158.0	68.0	6	农民	小	顺产	早?
10	6	20080006	39	158.0	66.5	8	工人	中	剖宫	早?
11	7	20080007	23	162.0	68.0	11	其他	小	剖宫	早?
12	8	20080008	20	162.0	70.5	4	管理	大	顺产	足?
13	9	20080009	27	160.0	71.5	3	其他	中	顺产	早?
14	10	20080010	22	162.0	70.0	7	工人	大	剖宫	足?
15	11	20080011	29	160.0	57.0	7	工人	中	顺产	足?
16	12	20080012	31	160.0	68.0	2	其他	文	顺产	早?
17	13	20080013	35	160.0	65.5	5	农民	小	顺产	早?
18	14	20080014	26	158.0	58.0	5	知识	大	顺产	足?

图 3-13　SAS 的 Output 窗口数据

② 单击“数据→数据工具→分列”，弹出“文本分列向导-第 1 步，共 3 步”（见图 3-14）。

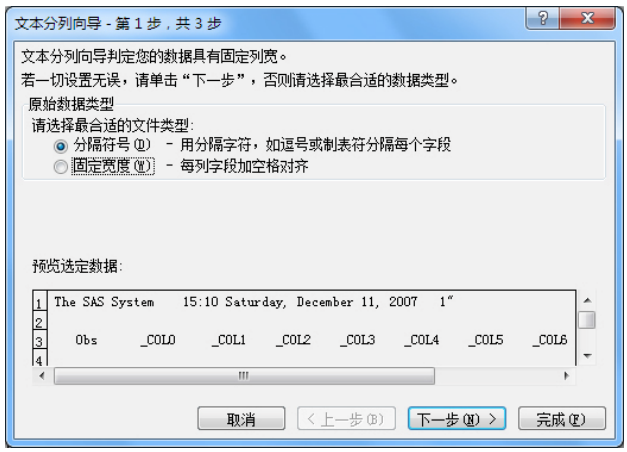


图 3-14　文本分列向导 1

③ 在该向导中给出了两种供选择的“原始数据类型”，即“分隔符号”和“固定宽度”。分隔符号有：Tab 键、分号、逗号、空格、其他，如果选择“其他”，则可在其后的方框中输入其他分隔符号，如“(”、“(”、“？”等（注意：中文符号与英文符号是有区别的）。本例选择“固定宽度”，单击“下一步”按钮，弹出“文本分列向导-第 2 步，共 3 步”；

④ 在该向导中提示设置字段（即变量）宽度。单击鼠标可建立分列线，双击鼠标可清除分列线，采用拖放方式可移动分列线（本例无须调整分列线）。完成分列线的设置后单击“下一步”，弹出“文本分列向导-第 3 步，共 3 步”；

⑤ 设置每列数据的类型（对于本例可忽略），单击“完成”；

⑥ 将不必要的行删除。如本例可删除第 1、2、4 行等。

如果选取图 3-14 中的“分隔符号”，单击“下一步”，会弹出如图 3-15 所示的界面。应该特别注意：

①其中的分号、逗号都是英文半角符号，如果是中文符号需要选取“其他”，并在“其他”后面的方框中键入相应的符号；

②在图 3-15 中的右上部有“连续分隔符号视为单个处理”选项，可视情况选用。

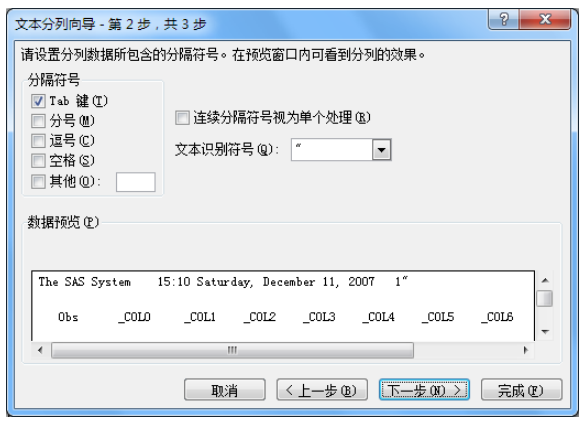


图 3-15　设置分隔符号



注意：数据分列，一次只能处理一列数据，否则，Excel 会给出错误提示信息。

3.7　添加与编辑批注

为了增加 Excel 建立的数据清单的通用性，有时有必要将变量名采用 8 个以内的英文字母表示；分类变量（如性别、血型）的取值也尽量采用数字表示，如性别以“1”表示“男”、“0”表示“女”，血型以“1”表示“A 型”、“2”表示“B 型”、“3”表示“AB 型”、“4”表示“O 型”。如果时间一长，这些变量名、具有特殊含义的数字、甚至产生数据公式的含义等，将会被遗忘。为了帮助用户自己或使用数据的统计人员记忆和理解相应单元格的信息，有必要对某一变量名或该变量的取值加以批注（comment）。

（1）添加批注

为了添加批注，首先选中需要添加批注的单元格，单击“审阅→批注→新建批注”，在弹出的批注框中输入批注内容，然后单击批注框外的工作表。此时，添加了批注的单元格右上角将出现一个小红点，表明该单元格已经被添加了批注。

（2）编辑批注

如果想修改其中的内容，选中要修改批注的单元格，“审阅→批注→编辑批注”，如图

3-16 所示；或者在加有批注的单元格单击鼠标右键，在弹出的快捷菜单中选择“编辑批注”命令。编辑批注方法与单元格编辑数据的方法相同。为了使其更清楚、美观，可以对字体、字号、颜色等作出修改编辑。

如果想删除批注，选中要修改批注的单元格，“审阅→批注→删除”；或者在加有批注的单元格单击鼠标右键，在弹出的快捷菜单中选择“删除批注”命令即可。



图 3-16　编辑批注

（3）查看批注

将鼠标指针移至包含批注（有小红点）的单元格上，该单元格的批注内容即可显示出来。

3.8　特殊数学符号的键入

统计学指标分为总体参数和样本统计量，一般情况下，总体参数采用希腊字母表示，比如总体均数 μ 、总体标准差 σ 、总体率 π；而样本统计量采用英文字母表示，比如样本均数 $\overline{x}$ 、样本标准差 s、样本率 p，具体参见表 3-1。

表 3-1　希腊字母与英文字母间的对应关系

编　号	希腊字母			英文字母	
	小　写	大　写	读　音	小　写	大　写
1	α（总体截距、Ⅰ类错误概率）	A	alpha	*a*	A
2	β（总体回归系数、Ⅱ类错误概率）	B	beta	*b*	B
3	χ（卡方符号）	X	chi	*c*	C
4	δ（总体处理组间的差别）	Δ	delta	*d*	D
5	ε（总体残差）	E	epsilon	*e*	E
6	ϕ（标准正态分布的面积值）	Φ（正态分布的面积值）	phi	*f*	F
7	φ		phi	*j*	
8	γ	Γ（伽玛函数）	gamma	*g*	G
9	η	H	eta	*h*	H
10	ι	I	iota	*i*	I
11	κ（总体一致性 Kappa 值）	K	kappa	*k*	K

续表

编　号	希腊字母			英文字母	
	小　写	大　写	读　音	小　写	大　写
12	λ	Λ	lamda	*l*	L
13	μ（总体均数）	M	mu	*m*	M
14	ν（自由度）	N	nu	*n*	N
15	*o*	O	omicron	*o*	O
16	π（总体率）	Π（连乘积符号）	pi	*p*	P
17	ϖ		pi	*v*	
18	θ	Θ	theta	*q*	Q
19	ϑ		theta	*J*	
20	ρ（总体相关系数）	P	Rho	*r*	R
21	σ（总体标准差）	Σ（求和符号）	sigma	*s*	S
22	ς		final sigma	*V*	
23	τ	T	tau	*t*	T
24	υ	Υ	upsilon	*u*	U
25	ω	Ω	omega	*w*	W
26	ξ	Ξ	xi	*x*	X
27	ψ	Ψ	psi	*y*	Y
28	ζ	Z	zeta	*Z*	Z

在表 3-1 中罗列出了 20 多个希腊字母的大小写、发音、统计学含义，以及相应的英文字母大小写。28 个小写希腊字母分别对应 26 个小写英文字母和 2 个大写英文字母（其中希腊字母ς对应于英文字母 V，希腊字母ϑ对应于英文字母 J）。

产生表 3-1 中的希腊字母可以采用 3 种方法，即采用公式编辑器、特殊符号、Symbol 字体。

（1）数学公式编辑器

微软 Office 软件（包括 Word、Excel、PowerPoint 等）中带有数学公式编辑器，基本上可以满足数学符号编辑的需要。调用方法是单击“插入→文本→对象”后在弹出对话框内下拉列表中选择“Office 公式 3.0”。也可单击“插入→符号→公式”。

比微软数学公式编辑器更好用的是 MathType 数学公式编辑器。在网上搜索关键词“MathType 数学公式编辑器”，可以找到免费下载软件“MathType 最新版本的数学公式编辑器”。本书及光盘文件含有很多公式，大多使用 MathType 编辑，如果您的光盘文件的显示和书上的不一样请注意在网上下载这类公式编辑器，比如可在下面网址 http://ishare.iask.sina.com.cn/f/13933028.html 的页面中下载 MathType 6.7。

（2）微软 Office 符号

微软 Office 软件（包括 Word、Excel、PowerPoint 等）中带有符号（或特殊符号）。使用时可单击“插入→符号→符号”，字体选“普通文本”，子集选“基本希腊语”，可获得所需要的符号。

（3）微软 Office Symbol 字体

在微软 Office 软件（包括 Word、Excel、PowerPoint 等）中，采用 Symbol 字体，均可将英文字母转变成希腊字母。例如，如果选中表 3-1（见光盘文件“03-06 特殊数学符号的键入.xlsx”）中的最后两列（即大、小写英文字母），将其字体修改为 Symbol 字体后，则倒数第二列与倒数第一列分别和该表的第二列（小写希腊字母）与第三列（大写希腊字母）一一对应，完全相同。



注意：

统计学符号采用正体还是斜体，具有一定的规范。例如公式、模型中代表某一变量意义的外文字母，一律使用斜体；代表其他意义的字母，如自然对数 e，对数符号 log、ln，三角函数符号 cos、sin，极大值、极小值符号 max、min，微分符号 d，增量符号Δ（请与三角形符号△区别），以及常量或者恒量等均用正体。正文及图表中外文字母的使用亦遵循此规则。

3.9 数据的导入与导出

Excel 可打开很多数据库文件，也可将 Excel 文件另存为其他数据库类型。

（1）打开其他数据库文件

Excel 可以打开其他类型的数据文件，如打开以 DBF 为扩展名的 dBase 数据库文件、Access 文件、文本文件等。具体操作方法如下：

①单击“文件”，选择“打开”，弹出“打开”对话框；

②在对话框右下侧框中选择所要打开的文件的类型（如 dBase 文件、Access 文件、文本文件）及其所在的目录；

③用鼠标单击要打开文件的文件名称，并按 Excel 提示进行操作即可。

（2）另存为其他类型的数据文件

Excel 文件同样也可另存为其他类型的数据文件，具体操作步骤如下：

①编辑好文件后，选择“文件→另存为”，弹出“另存为”对话框；

②在“另存为”对话框中选择所要打开文件的类型及其所在的目录；

③输入文件名之后，单击“保存”按钮即可。

改进的 Excel 2010 版本，可以存为 pdf（Portable Document Format，便携文件格式）、ods（Operational Data Store，操作型数据存储）等文件。



注意：为了利用统计软件进行数据分析，Excel 数据清单应整理为图 3-7 的形式。即第 1 行为变量名称，每一列代表一个变量；每一列的变量名采用 8 个以内英文字母表示，如果变量名是中文，某些统计软件程序将不运行（或变成乱码），并提示变量名出错；有的可能用其他名称代替。如果 Excel 文件的某些单元格中没有任何数据（空的单元格），转换到统计软件之后，统计软件将它们默认为缺失数据，常用“.”表示。

第 4 章

数据的编辑与查询

数据清单建立后，可能需要继续对部分数据进行剪切、复制、移动、删除、替换、产生新数据。修改数据清单的目的就是要利用这些数据，从中查询、提取所需的信息，帮助科学管理或科学研究。本章先介绍一些数据编辑的方法，然后介绍如何利用数据清单进行排序、筛选等，以便完成数据查询与新数据清单建立等工作。

4.1　窗口的冻结与拆分

对于较大的数据清单，由于工作表过宽过长，因此，在窗口中查看、编辑整理数据时，往往出现顾头顾不了尾的现象。尤其当前单元格处于工作表右下部时，常常搞不清当前单元格属于工作表的哪一个记录（行，观测个体）的哪一个字段（列，变量），为了解决此问题可采用“视图”选项卡，“窗口”命令组中的“冻结窗格”或“拆分”按钮。

（1）冻结窗格

采用“冻结窗格”后，上窗格、左窗格被冻结。通常用冻结列字段名（如住院号、年龄、身高、体重、住院天数等，上窗格）和行记录编号（如住院号 2025655、2025653、2025830、…，左窗格），使记录编号和变量名始终出现在屏幕上，然后通过移动水平或垂直滚动条，来查看工作表的其他部分内容。

“冻结窗格”的操作步骤如下：

① 选取作为冻结点的单元格，如打开光盘文件“04-01 妊娠分娩.xlsx”，选取单元格 B2（见图 4-1）。

	A	B	C	D	E	F	G	H	I	J
1	住院号	年龄（y）	身高（cm）	体重（kg）	住院天数	职业	文化程度	分娩方式	妊娠结局	
2	20080001	25	162	76.0	9	其他	中学	顺产	早产	
3	20080002	32	153	60.0	7	其他	小学	剖宫产	足月	
4	20080003	28	158	64.0	10	其他	中学	顺产	足月	
5	20080004	29	162	68.0	8	工人	大学	剖宫产	足月	
6	20080005	27	158	68.0	6	农民	小学	顺产	早产	
7	20080006	39	158	66.5	8	工人	中学	剖宫产	早产	
8	20080007	23	162	68.0	11	其他	小学	剖宫产	早产	
9	20080008	20	162	70.5	4	管理人员	大学	顺产	足月	
10	20080009	27	160	71.5	3	其他	中学	顺产	早产	
11	20080010	22	163	70.0	7	工人	大学	剖宫产	足月	

妊娠分娩

图 4-1　冻结窗格的冻结点选取

② 单击“视图→窗口→冻结窗格”（见光盘文件“04-01 妊娠分娩.xlsx”）。

完成冻结窗格后，冻结点单元格以上或左侧的所有单元格将被冻结，不管鼠标向哪个方向移动，它们始终保留在屏幕上。

当不需要冻结时，单击“窗口→撤销窗口冻结”即可。

（2）拆分

“拆分”与“冻结窗格”的作用基本相似。其区别在于：“冻结窗格”后，只能滚动查看右下窗格的工作表内容；而“拆分”后，每个被拆分的窗格都可以通过移动滚动条来显示工作表内容。具体操作步骤基本相同，比如对于同一数据清单，选定单元格 B2 后，单击命令组的“窗口→拆分”，向下、向右分别移动垂直、水平滚动条若干距离，可获得如图 4-2 所示的效果。

	A	B	C	D	E	F	G	H	I	J
1	住院号	年龄（y）	身高（cm）	体重（kg）	住院天数	职业	文化程度	分娩方式	妊娠结局	
2	20080001	25	162	76.0	9	其他	中学	顺产	早产	
3	20080002	32	153	60.0	7	其他	小学	剖宫产	足月	
4	20080003	28	158	64.0	10	其他	中学	顺产	足月	
5	20080004	29	162	68.0	8	工人	大学	剖宫产	足月	
6	20080005	27	158	68.0	6	农民	小学	顺产	早产	
7	20080006	39	158	66.5	8	工人	中学	剖宫产	早产	
8	20080007	23	162	68.0	11	其他	小学	剖宫产	早产	
9	20080008	20	162	70.5	4	管理人员	大学	顺产	足月	
10	20080009	27	160	71.5	3	其他	中学	顺产	早产	

图 4-2　拆分后的查看效果

4.2　数据的剪贴与删除

1. 数据的插入与删除

当建立工作表格、输入编辑整理数据时，经常需要插入或删除单元格、整行、整列。

（1）单元格区域的插入

在选取的单元格或单元格区域，单击鼠标右键，会出现一个快捷菜单，单击功能区中的“插入”，屏幕上会出现一个对话框，有 4 个选项：活动单元格右移、活动单元格下移、整行、整列（见图 4-3 左），可根据需要进行选取，单击“确定”完成插入。

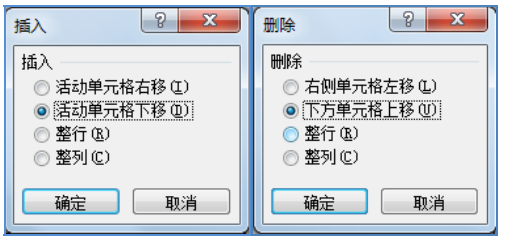


图 4-3　插入（左）与删除（右）的选项

（2）单元格区域的删除

在选取的单元格或单元格区域，单击鼠标右键，会出现一个快捷菜单，单击快捷菜单中的“删除”，屏幕上会出现一个对话框，也有 4 个选项：右侧单元格左移、下方单元格上移、整行、整列（见图 4-3 右），可根据需要进行选取，单击“确定”完成删除。

（3）整行或整列的插入

在选取的行号上单击鼠标右键，会出现一个快捷菜单，单击快捷菜单中的“插入”，则在当前行之上插入一行。即当前行下移一行。

在选取的列标上单击鼠标右键，会出现一个快捷菜单，单击快捷菜单中的“插入”，则

在当前列的左侧插入一列。即当前列右移一列。

（4）整行或整列的删除

在选取的行号（或列标）上单击鼠标右键，会出现一个快捷菜单，单击快捷菜单中的“删除”，则该行（或列）被删除，下面各行向上（或右侧各列向左）移动补充删除的行或列。如果不小心删除了不该删除的区域或整行、整列，可以通过单击快速访问栏中的“撤销”图标或使用快捷键“Ctrl+Z”来恢复被删除的内容。“撤销”操作是 Excel 中较常用到的操作，如果不小心实施了错误的操作，那么可以通过“撤销”操作使工作表内容恢复原样。

2. 数据的移动和复制

（1）整列或整行的移动

采用“剪切（Ctrl+X）”较方便。在需要移动的列标（或行号）处单击鼠标右键，会出现一个快捷菜单，单击快捷菜单中的“剪切”；再在需要插入两列之间的右一列列标上（或两行之间的下一行行号上）单击鼠标右键，选择快捷菜单中的“插入已剪切的单元格”，则“剪切”数据可移至需要移动的位置。

（2）单元格或单元格区域的移动

采用鼠标拖放较方便。用鼠标选取所要移动的单元格或单元格区域，将鼠标指针移到单元格或单元格区域的边框，此时鼠标的空心十字变为十字箭头与箭头混合的图标，按住鼠标左键后变为箭头，将选取的单元格或单元格区域拖放移动至所需要的地方。

（3）复制（Ctrl+C）

采用鼠标右键较方便。首先选取要复制的单元格或单元格区域（或整行、整列），单击鼠标右键，在出现的快捷菜单中单击“复制”命令；将鼠标指针指向要存放数据的单元格，单击鼠标右键，这时会弹出快捷菜单，单击该菜单中的“粘贴(Ctrl+V)”，即可完成复制。

3. 数据的清除

在选取了单元格或单元格区域的条件下，采用 Delete 键、BackSpace 键或空格键可清除单元格或单元格区域中的数据。

也可在选取的单元格或单元格区域处单击鼠标右键，在出现的快捷菜单中单击“清除内容”命令。

4.3　数据的查找与替换

数据编辑整理时，需要利用“查找与替换”将中文文字数字化，如将光盘文件“04-01 妊娠分娩.xlsx”的文化程度“小学”替换为“1”、“中学”替换为“2”、“大学”替换为“3”。

此外，如果数据清单中出现错误，需要成批修改错误也要用到“查找与替换”功能。

选取整个工作表、整行、整列或单元格区域，选择“开始→编辑→查找和选择”，或按下“Ctrl+H”组合键，在弹出的“查找和替换”对话框中，填写“查找内容”和“替换为”，单击“查找下一个”、“替换”或“全部替换”按钮，可完成数据替换任务。

如果查找和替换需要区分大小写、区分全/半角等，可以单击“查找和替换”对话框中的“选项”按钮，选中所需更改的选项（见图 4-4）。

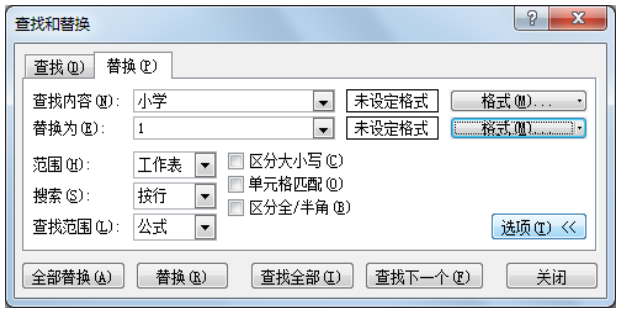


图 4-4　扩展后的“查找和替换”对话框

注意：在“查找内容”下拉列表框中可以使用通配符“*”或“？”。通配符“？”代表单个任意字符，而通配符“*”代表一个或多个任意字符。查找内容“小*”与“小学”等价。此外，按住 Shift 键，然后单击“查找下一个”按钮，可以返回查找前一个符合条件的内容。

完成替换后，Excel 会出现提示信息框，如图 4-5 所示。由此可检查数据查找和替换是否正确，单击“确定”。如果替换错误，可在工具栏中单击“撤销”图标来恢复被不正确替代的内容。

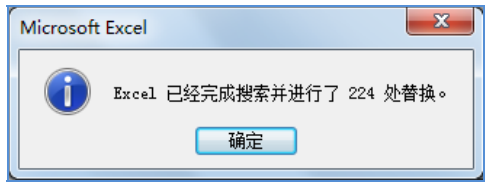


图 4-5　查找和替换提示信息框

4.4　行列转置与选择性粘贴的其他妙用

在 Excel 中，当复制了某单元格或单元格区域的数据后，粘贴数据有两种方式，即粘

贴与选择性粘贴。选择性粘贴有很多选项，如粘贴、运算、跳过空单元（格）、转置等（见图 4-6），如果能灵活运用选择性粘贴的这些功能，对于整理数据将有很大帮助，在 Excel 2010 中选择性粘贴功能进行了改进，很多“选择性粘贴”选项卡上的常用功能如“转置”、“格式粘贴”，“值粘贴”等添加到了“粘贴”功能下，提高了使用的便捷性，如图 4-6（左）所示。

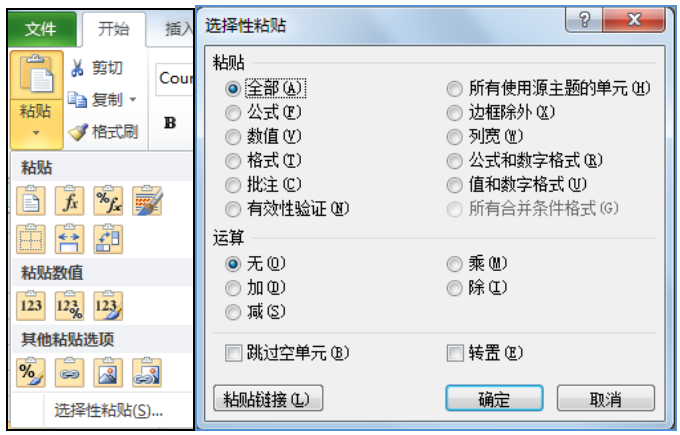


图 4-6　数据粘贴选项（左）与选择性粘贴选项卡（右）

1. 转置——数据的行列互换

矩阵运算时需要计算转置矩阵，其他统计计算有时也需要将数据的行列互换。行列互换在 Excel 中相当容易，只需选取需要行列互换的单元格区域，单击“复制”按钮，然后选择“开始→剪贴板→粘贴▼→转置（T）”，或在“剪贴板→粘贴▼→选择性粘贴”打开“选择性粘贴”对话框，在右下角选中“转置”，单击“确定”即可。

原先的第 1 行转置后变成了第 1 列，原先的第 2 行转置后变成了第 2 列依次类推。将图 4-1 中数据（见光盘文件“04-01 妊娠分娩.xlsx”）的前 10 行转置后的结果如图 4-7 所示。因为住院号的数字位数较长，所以显示为“#######”，如果增加每一列的宽度，住院号显示将恢复正常。

	A	B	C	D	E	F	G	H	I	J
1	########	########	########	########	########	########	########	########	########	########
2	25	32	28	29	27	39	23	20	27	22
3	162	153	158	162	158	158	162	162	160	162
4	76.0	60.0	64.0	68.0	68.0	66.5	68.0	70.5	71.5	70.0
5	9	7	10	8	6	8	11	4	3	7
6	其他	其他	其他	工人	农民	工人	其他	管理人员	其他	工人
7	中学	小学	中学	大学	小学	中学	小学	大学	中学	大学
8	顺产	剖宫产	顺产	剖宫产	顺产	剖宫产	剖宫产	顺产	顺产	剖宫产
9	早产	足月	足月	足月	早产	早产	早产	足月	早产	足月

图 4-7　图 4-1 中数据经转置后的结果

2. 数值的选择性粘贴

一般的粘贴是将复制的单元格所有信息（包括公式、数值、格式与批注等）全部粘贴到另一个单元格。很多情况下并不需要如此，例如，为了利用由公式计算的数据，复制该列数据，然后粘贴到其他单元格后，由于公式的相对引用等原因，粘贴的数据将可能变成其他数据值，而非先前复制的数据。解决这一问题的方法是：单击“开始→剪贴板→粘贴▼→数值”，即可将先前公式计算的数据复制到所需粘贴的位置。

有时，为了让公式计算出的数据移动位置后，其数值保持不变。可先选取该单元格区域，复制后选择性粘贴“数值”到原单元格（左上角）位置，此时公式等不复存在，该列变成了纯“数据”，可以任意移动位置而保持数值不变。

3. 选择性粘贴的运算选项

选择性粘贴的“运算”有加、减、乘、除等选项（见图 4-6 右）。有时需要将某数据加、减、乘、除到某一列的每个个体中。例如，为了将身高的单位“米”转换成“厘米”，需要将身高乘以 100，可按下列操作进行：

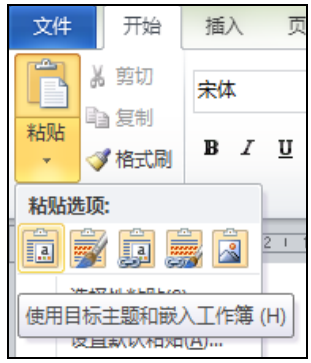


图 4-8　在 Word 中选择性粘贴 Excel 图形对话框

首先在某个空白单元格中键入 100，选定此单元格，单击“复制”按钮，选取想修改的单元格区域，例如 E2:E101；然后在“开始→剪贴板→粘贴▼→选择性粘贴→运算”选项中单选“乘”，单击“确定”。为了保持数据清单的整洁，可将某个空白单元格中输入的 100 删除。

4. Excel 统计图粘贴到 Word

将 Excel 统计图形，按照一般的复制、粘贴方法复制到 Microsoft Word 2003 及前期版本中后，统计图的字体、坐标刻度有时会发生不期望的自动改变。Office 2010 的有了较大中，当复制完 Excel 统计图形后，在 Word 中单击“开始→剪贴板→粘贴▼”会出现如图 4-8 所示的 5 个粘贴选项（后单击鼠标右键在弹出快捷选项条中出现 5 个粘贴选项）。这 5

个选项分别是“使用目标主题和嵌入工作簿”、“保留源格式和嵌入工作簿”、“使用目标主题和链接数据”、“保留源格式和链接数据”、“图片”。最好选择前面两个，即“使用目标主题和嵌入工作簿”与“保留源格式和嵌入工作簿”，因为直接将 Excel 数据同时也带入到了 Word（或 PPT 等）文件中，便于以后编辑修改这些 Excel 图形。

4.5　将连续型变量划分等级

用 IF 函数可以实现连续型变量的等级划分。例如，当统计学生成绩时，希望 60 分以下的分数显示为“不及格”，60 分以上的分数显示为“及格”。假如分数存放在 C3 单元格中，要将等级划分结果显示在 D3 单元格中，那么在 D3 单元格输入公式“=IF（C3<60,"不及格","及格"）”可完成以上任务。

IF 函数有三个参数，第一个参数为“条件”，第二个参数为“条件”成立时的结果，第三个参数为“条件”不成立时的结果。IF 函数中可多次嵌套 IF 函数，Excel 2003 最多可嵌套 7 层，Excel 2007 及以上版本可嵌套 64 层。

下面举一个嵌套 3 层 IF 函数的例子。如将学生成绩按小于 60 分、60～70 分（不包括 70 分）、70～90 分（不包括 90 分）、90 分及以上分别划分等级为不及格、及格、良好、优秀。假如分数存放在 C3 单元格中，要将等级划分结果显示在 D3 单元格中。那么在 D3 单元格输入公式“=IF(C3<60,"不及格",IF(C3<70,"及格",IF(C3<90,"良好","优秀")))”可完成任务。

如果第 C 列中有一个小班的 19 名学生的成绩，数据存放在光盘文件“04-02 学生成绩等级划分.xlsx”的单元格区域 C3:C21 中，为了获得每名学生的分类结果，可先单击单元格 D3，将鼠标移至 D3 右下角填充柄处，向下拖放公式至单元格 D21，可完成整个小班的成绩等级划分（见图 4-9）。

D3　fx　=IF(C3<60,"不及格", IF(C3<70,"及格", IF(C3<90, "良好", "优秀")))

	A	B	C	D	E	F	G
1	xx年级学生数学成绩						
2	学号	姓名	成绩	等级分			
3	894510	张1	78	良好			
4	894511	刘1	73	良好			
5	894512	马1	82	良好			
6	894513	张2	67	及格			
7	894514	刘2	57	不及格			
8	894515	李1	87	良好			
9	894516	倪1	91	优秀			
10	894517	柳1	74	良好			
11	894518	李2	83	良好			

图 4-9　将成绩划分为等级

4.6 年龄的计算与当前日期的生成

Excel 有一个函数 DATEDIF()可以计算两个日期之间的年、月或日数，获得的值为整数，可用此公式计算实足的年龄、月龄、日龄。其公式格式为：

```
DATEDIF(日期 1,日期 2,"y")
```

其中日期 1 为先前（过去）的日期，日期 2 为当前的日期，如果日期为 1991 年 2 月 28 日，则日期可以按“1991-2-28”形式输入。第三个参数采用英文引号（" "）引起来，引号内的 y 表示 year（年），该引号中也可用 m（month，月）、d（day，天）。

例 4-1：在单元格 A1 存放某大四学生的出生日期“1991-2-28”，在单元格 B1 存放当前日期“2013-1-27”，计算该大四学生的年龄、月龄与日龄，并将其结果存放在单元格 C1 中。

解：单击单元格 C1，在编辑栏输入公式“=DATEDIF(A1,B1,"y")”，按 Enter 键，在单元格 C1 中得到 21，即该大四学生的实足年龄（而非虚岁）为 21 岁。如果将公式中的 y 分别改为 m、d，则可得该高中生的实足月龄、实足日龄分别为：262、8004。

注意：计算获得的结果是整数，且为实足的年、月或日数；为了获得出生日期至“今天”（一个可变日期）的年龄，还可利用公式“=DATEDIF(A1,TODAY(),"y")”计算。

如果“日期 1”为入院日期，“日期 2”为出院日期，y 改为 d，可以用此公式计算医院住院病人的住院天数。

按如下方法可插入当前时间和日期：

①插入当前日期：按下 Ctrl 键后不放，再按“;”键。如：2013-1-27。

②插入当前时间：同时按下 Ctrl 键和 Shift 键后不放，再按“;”键。如：21:14。

③插入当前日期和时间：按下 Ctrl 键后不放，再按“;”键；继续键入一个空格，同时按下 Ctrl 键和 Shift 键后不放，再按“;”键。得到结果如：2013-1-27 21:15。

4.7 数据的跨工作表操作

Excel 的一个数据文件就是一份工作簿，在每一份工作簿中可存放 255 张工作表，工作表之间的数据可以协作作业完成许多琐碎而又相关的工作。

1. 工作表间数据的引用

例 4-2：假设在同一份工作簿中有名称为 Sheet1、Sheet2 和 Sheet3 的三张工作表，现要用 Sheet1 的 D8 单元格的内容乘以 20%，再加上 Sheet2 的 B8 单元格内容乘以 50%作为 Sheet3 的 A8 单元格的内容。写出 Excel 计算公式。

解：在 Sheet3 的 A8 单元格输入公式“=Sheet1!$D$8*20%＋Sheet2!$B$8*50%”可简单完成跨工作表计算。公式中工作表名称（如 Sheet1）与单元格名称（如$D$8）之间采用感叹号“!”分隔。采用鼠标选取工作表，再选取单元格方式可更准确地建立公式。单元格的引用采用绝对引用，可以使公式移动到其他单元格后，所引用的单元格不会发生变化。

2. 工作簿间数据的引用

当需要引用不同工作簿间的数据时，可在工作表间数据引用的基础上，加上工作簿名称并用括号“[]”括起来。引用格式为：

[工作簿名称]工作表名称！单元格名称

比如引用“练习数据”工作簿中“Sheet1”工作表的 D8 单元格，可表示为“[练习数据]Sheet1!$D$8”。如果引用工作簿未打开，则在引用时应写出该工作簿存放位置的绝对路径，并用单引号括起来。例如：“’e:\Excel 与数据分析\[练习数据]Sheet1’!$D$8”。

3. 三维引用

当一个工作簿中多个工作表在相同单元格或单元格区域的数据性质相同时，如均为各医院科室月门诊量，可采用三维引用计算。

例 4-3：某医院“门诊量”工作簿中存放了 12 张“各科室月门诊量”工作表“门诊量 1”～“门诊量 12”，每张工作表的格式相同，每张工作表的单元格 F3:F6 分别存放内、外、妇、儿月门诊人次，求这一医院一年里内、外、妇、儿门诊量总人次。

解：采用公式“=SUM(门诊量 1:门诊量 12！$F$3:$F$6)”可简单计算获得该医院 1 年中内、外、妇、儿门诊总人次。

除“求和”SUM 函数外，也可采用 Excel 的 AVERAGE、PRODUCT、MAX、MIN、COUNT、COUNTA、VAR、VARP、STDEV、STDEVP 函数等分别实现三维引用的求均数（即平均值）、连乘（即乘积）、最大值、最小值、非空白单元格的计数（即计数）、数字单元格的计数（即数值计数）、方差、总体方差、标准差（即标准偏差）、总体标准差（即总体标准偏差）等。

4. 同时编辑多个工作簿文件

一般情况下，窗口排列采用“平铺”方式，不便于同时查看或编辑多个工作簿文件。

为了同时查看或编辑多个工作簿，可单击“视图→窗口→全部重排”，在“重排窗口”对话框中选择“水平并排”或“垂直并排”。如果要同时显示活动工作簿中的工作表，则选中“当前活动工作簿的窗口”。如果要将工作簿还原成整屏显示，则单击工作簿窗口右上角的最大化按钮。

5. 工作簿间的数据传递

工作表的移动和复制不仅可以在同一工作簿中进行，而且还可以在不同的工作簿间进行，两者的操作基本相同。如果要移动工作表，可采用鼠标拖动工作表名称到同一工作簿或其他工作簿的指定位置，然后松开鼠标按键；如果要复制工作表，应先按住 Ctrl 键，再采用鼠标拖动工作表名称到同一工作簿或其他工作簿的指定位置，然后松开鼠标按键。当然也可以使用编辑菜单中的移动或复制工作表命令实现工作表的移动与复制。

4.8 多个工作表的合并计算

利用 Excel 的“合并计算”功能，可以将多个工作表的数据合并计算结果存放到另一个工作表中。比上一节采用三维引用函数（共 11 个函数，见例 4-3）计算更加方便、快捷。

在合并计算中，存放合并计算结果的工作表称为“目标工作表”，其中存放合并计算结果的区域称为目标区域，目标工作表应该是当前工作表，目标区域也应该是当前单元格区域。而被合并计算的各个工作表称为“源工作表”，其中被合并计算的数据区域称为来源区域，源工作表可以是打开状态，也可以是关闭状态。

Excel 提供了两种合并计算数据的方法：按位置合并计算和按类合并计算。

1. 按位置合并计算

按位置合并计算既常用也简单。不过，这一计算方法要求被合并的各工作表格式必须相同。这种资料在日常工作中经常碰到，例如，建立某公司每个季度或每月报表的年终汇总；某省市各下属单位年终传染病发病情况报表汇总；某医院各科室月门诊人次年终汇总等。

例 4-4：某市 6 所医院 1983-2002 年的出院人数、出院病人数、出院者占用总床日数的数据分别放在工作表名称为医院 1、医院 2……医院 6 的 6 张工作表（6 张工作表的结构相同，其中“医院 1”的数据见图 4-10，也见光盘文件“04-03 多个工作表的合并计算.xlsx”）中。试计算这 6 所医院各年度三个指标的和。

解：合并计算 6 个工作表的具体操作步骤如下：

① 复制“医院 1”工作表，将复制的工作表名称改为“合并计算”，选择单元格区域 B2:D21，按 Delete 键，删除数据保留单元格格式。

② 选定 B2:D21 为目标区域。



	A	B	C	D
1	年份	出院人数	出院病人数	出院者占用总床日数
2	2002	4529	3921	60328
3	2001	1732	1542	35379
4	2000	1856	1648	34331
5	1999	1824	1581	36311
6	1998	2105	1837	53829
7	1997	2124	1866	60748
8	1996	2441	2173	68239
9	1995	2032	1836	58415
10	1994	1950	1847	63988
11	1993	1981	1905	57309
12	1992	1966	1966	55334
13	1991	1977	1977	50500
14	1990	1699	1699	45871
15	1989	1699	1699	45045
16	1988	1706	1706	45181
17	1987	1990	1990	47643
18	1986	1804	1804	41742
19	1985	2267	2267	37699
20	1984	1082	1082	24753
21	1983	1599	1599	32915
22				
23				

医院1 医院2 医院3 医院4 医院5 医院6

图 4-10　6 所医院中“医院 1”的数据

③单击“数据→数据工具→合并计算”，弹出“合并计算”对话框（见图 4-11）。

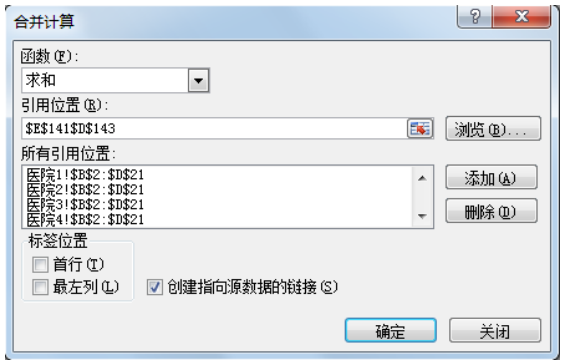


图 4-11　“合并计算”对话框

④在“函数”下拉列表框中选“求和”函数【这里一共可选求和、均数（即平均值）、连乘（即乘积）、最大值、最小值、非空白单元格的计数（即计数）、数字单元格的计数（即数值计数）、方差、总体方差、标准差（即标准偏差）、总体标准差（即总体标准偏差）11 个函数】。

⑤在引用位置框中，定义要合并的来源区域，先用鼠标单击要合并计算的源工作表名称“医院 1”，采用鼠标拖放方式选取来源区域 B2:D21。

⑥选取一个引用位置后，单击“添加”，则“引用位置”框中的内容增加到“所有引用位置”框中。

⑦重复上述⑤、⑥两步，直到所有要合并计算的来源区域都添加完毕。

⑧为了保证合并结果与源数据自动保持一致，选定“创建连至源数据的链接”。单击“确定”后，获得的结果如图 4-12 所示。

	A	B	C	D
1	年份	出院人数	出院病人数	出院者占用总床日数
8	2002	67046	60502	1072211
15	2001	60283	54626	1006046
22	2000	60946	56958	1019741
29	1999	55342	51989	1012935
36	1998	53643	49999	1023496
43	1997	52556	48768	1045615
50	1996	50775	46573	1045758
57	1995	50560	45865	1047083
64	1994	45769	40785	906654
71	1993	44555	39063	854501
78	1992	42956	37078	788312
85	1991	46134	39434	829056
92	1990	47876	39766	845808
99	1989	36777	36777	734805
106	1988	35391	35391	697715
113	1987	36724	36724	702724
120	1986	32781	32781	671042
127	1985	33244	33244	657013
134	1984	33361	33361	670090
141	1983	35097	35097	673640

图 4-12　合并计算的结果

由图 4-12 可见，工作表的左侧出现了分级显示按钮。如果希望查看某年份合并数据的细节，可单击该年数据左侧的 ✚ 按钮，此时可见：该年份 6 所医院三个指标的数值及其合并计算值；如果希望查看全部数据细节，可以单击按钮 2，此时可见：1983-2002 年 6 所医院三个指标的数值及其合并计算值；如果再单击按钮 1，将返回到图 4-12 状态（见光盘文件“04-04 多个工作表的合并计算结果.xlsx”）。

注意：如果合并计算时，未选定“创建连至源数据的链接”，则合并的工作表中将不出现分级显示按钮，合并计算的结果是一系列常数；当来源区域的数据变动时，合并计算的结果不变，为此，还需要重新进行合并计算。如果选定了“创建连至源数据的链接”，则合并计算的工作表中是相应的公式，这样，当来源区域的数据变动时，合并计算的结果能自动更新，从而保证了源工作表与合并计算工作表的一致性。

2. 按分类合并计算

如果几个被合并计算的工作表结构不相同，则不能采用上述按位置合并计算的方法，而应根据分类合并工作表。该方法与上述按位置合并计算的方法的不同之处在于：选择目标区域时，应包括分类标志所在的行或列（即包括变量名），并在左下方“标志位置”处根据分类标志选择“首行”或“最左列”。

4.9　查询与新建子数据清单

为了把符合条件的个体（或记录）从数据清单中查找并筛选出来，当样本含量较少时可直接在数据文件中查找；但如果样本含量很大，且数据排列杂乱无章时，就需要采用排序和筛选的方法进行查询。

此外，求百分位数、求排列秩次（排位）、分类汇总、绘制光滑曲线图等也需要对数据按大小顺序排列。排序和筛选操作是一般数据库管理软件都具备的功能，但 Excel 的排序和筛选操作更加方便、直观。

4.9.1　排序

所谓排序（sorting）是指以当前数据清单中的一个或几个变量（字段）为关键字，对整个数据清单的所有个体进行重新排序。排序可以按升序，也可以按降序进行。对于数字型变量，按数值的大小顺序排序；对于字符型变量，按 ASCII 码大小排序；对于日期型变量，按日期的先后排序；中文字按拼音或笔画排序。

1. 升、降序按钮排序

这种排序方式非常快捷、方便。先用鼠标指针单击一个排序“关键字”，如“住院号”变量名（注意：不能错误单击该变量的列标，否则将只对该列数据排序），然后单击“数据→排序和筛选→升序”，即可对整个数据清单按指定变量名“住院号”按升序排序。

如果单击的是“降序”命令，则按降序进行排序。

2. 多个“关键字”排序

使用数据的升、降序按钮排序操作虽然简单，但不能满足复杂的排序要求，只能按单个变量的内容进行排序。如果需要对“04-01 妊娠分娩.xlsx”文件按多个“关键字”排序，如按“年龄”由大到小、“身高”由低到高、“职业”按字母排序，这时无法采用升、降序按钮排序，而应采用“数据→排序和筛选→排序”命令实现排序，具体操作步骤如下：

① 选定数据清单中任意单元格为当前单元格。

② 单击“数据→排序和筛选→排序”，弹出“排序”对话框。

- 在“排序”对话框中指定“主要关键字”为“年龄”，排序方式为“降序”；“次要关键字”为“身高”，排序方式为“升序”；“第三关键字”为“职业”，排序方式为“字母顺序”（可由右上角的“选项”中获得）（见图 4-13）。单击“确定”，可完成排序。

Excel2003“排序”对话框每次最多只对按 3 个变量进行多个“关键字”排序，而 Excel2007“排序”对话框每次最多可以按 64 个变量进行多个“关键字”排序。整个数据清单先按主要关键字排列，关键字相同者排在一起；若指定了次要关键字，则主要关键字相同者再按次要关键字排列；若指定了第三关键字，则主要关键字、次要关键字均相同者再按第三关

键字排列，……。

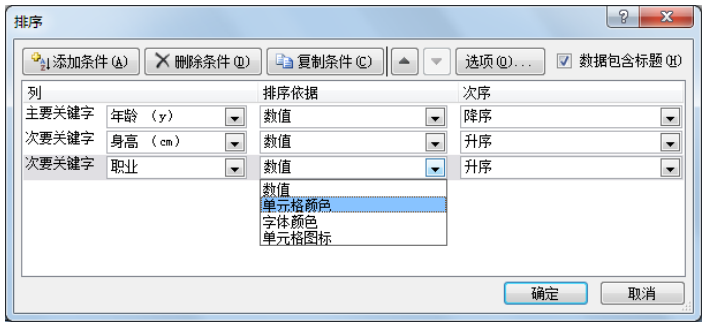


图 4-13　多个“关键字”排序对话框

注意：

①在排序对话框的右上方“数据包括标题”的选项。计算机一般会自动识别，但有时也需要人工调整。如果你的数据有变量名称（即字段名）存在，则需要选“数据包括标题”（见图 4-13）。

②图 4-13 右上方的“选项”中有区分大小写、方向（按列或行排序）、方法（字母或笔画）的备选项，可根据需要使用。

③在 Excel 2010 版本中在排序依据中又加入了“单元格颜色”“字体颜色”“单元格图标”三项排序依据，使得排序方式更加多样便捷。

4.9.2　自动筛选

从众多的数据中挑选出一部分满足条件的数据，可利用“筛选（filtering）”方法。选取筛选出的数据，然后复制、粘贴，可以此建立一个新的数据文件，供进一步数据分析用。

Excel 的筛选主要分为自动筛选和高级筛选。

自动筛选（autofilter）是一种快速的筛选方法，它可以方便地将满足条件的个体显示在工作表上，而不满足条件的个体被隐藏起来。自动筛选步骤如下：

①在工作表数据清单中选择任意一个单元格（如 H8）为当前单元格。

②单击“数据→排序和筛选→筛选”（如光盘文件“04-05 自动筛选.xlsx”）。该命令是一个选项开关，默认为关闭状态，当处于打开状态时，在数据清单中的每一列字段名旁边都会出现一个下拉箭头按钮。

③单击某一个下拉箭头按钮，在列表框中选定所需筛选的条件。条件选项有：升序排列、降序排列、按颜色排序、文本/数字筛选、搜索框、复选框等（见图 4-14）。如图 4-14 所示，职业的复选框具体值显示为工人、管理人员、农民、其他、商业服务、知识分子。

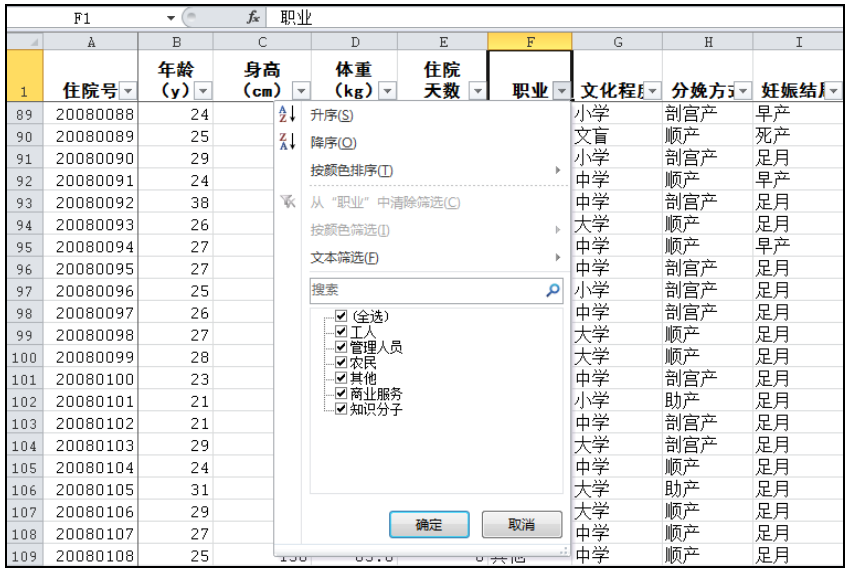


图 4-14　自动筛选

④使用以下操作，Excel 将按用户所给条件自动筛选出符合要求的记录。

a．使用“搜索”框输入要搜索的文本或数字；

b．通过单击复选框来清除或打对钩，去除或选取所要显示的数据内容；

c．使用高级条件查找满足特定条件的值；

在工作表中，满足条件的记录所在的行号和设置了筛选条件的变量名旁筛选箭头都呈现为蓝色，以提醒用户注意当前筛选的结果和所筛选变量。

此外，在工作表的左下方的状态栏给出了筛选的总结信息。如查找“04-05 自动筛选.xlsx”清单中满足职业为“工人”、妊娠结局为“死产”的所有记录，此时在工作表左下角状态栏显示“在 1402 条记录中找到 6 个”（见图 4-15）。

单击数值型数据的筛选下箭头，会有“数字筛选”，单击字符型数据的筛选下箭头，会有“文本筛选”（见图 4-14）。“文本筛选”有“等于”、“不等于”、“开头是”、“结尾是”“包含”、不包含、“自定义筛选”等选项，而“数字筛选”有“等于”、“不等于”、“大于”、“大于或等于”、“小于”、“小于或等于”、“介于”、“10 个最大的值”、“高于平均值”、“低于平均值”、“自定义筛选”等选项。“10 个最大的值”选项只对数字型变量有效，可以筛选最大值，也可筛选最小值；可以筛选 10 个，也可以自行设定筛选的数量；此外还可以设定“项”与“百分比”（见图 4-16）。



图 4-15　筛选的状态栏总结信息

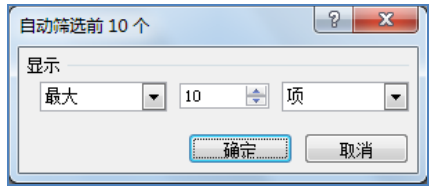


图 4-16　“自动筛选前 10 个”对话框

对于较为复杂的筛选条件，可以使用“自定义筛选”。例如，要筛选年龄大于等于 20，小于 40 的个体，其操作步骤如下：

单击年龄变量旁的筛选箭头，选择筛选列表中的“自定义筛选”选项，这时弹出“自定义自动筛选方式”对话框（见图 4-17）。

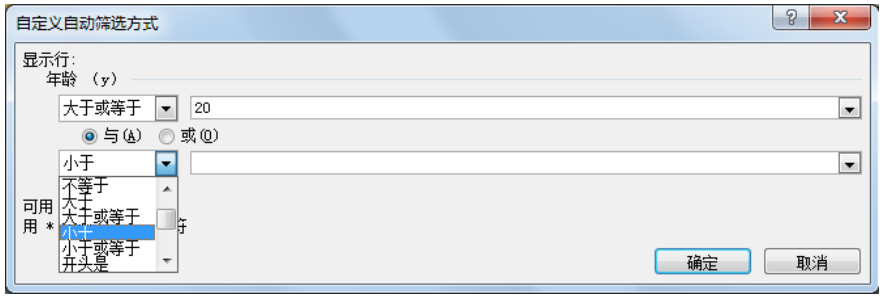


图 4-17　“自定义自动筛选方式”对话框

在“自定义自动筛选方式”对话框中，最多可设置两组筛选条件。每一组筛选条件由关系运算符构成条件表达式，其中左侧运算符有等于（equals）、不等于（not equal）、大于（greater than）、大于或等于（greater than or equal to）、小于（less than）、小于或等于（less than or equal to）、始于（begins with）、止于（ends with）、并非起始于（not begin with）、并非结束在（not end with）、包含（contains）和不包含（not contain）等。在对话框右侧，可以单击下拉框箭头选择值，也可以直接键入值。两个条件间的关系可以是“和”（两者都满足），也可以是“或”（其中之一满足）。

为了利用筛选后的数据结果，可以先选中筛选出的数据清单，再复制、粘贴。单击工作表的左上角（行号列标交叉处）选中全部筛选数据，按住 Ctrl 键选中多个单元格区域数据，等等；然后将复制的内容粘贴到一张新的工作表中。

若要取消某个变量的筛选条件，可单击该变量右侧的筛选箭头，在下拉框中选择“全选”作为筛选条件即可。当完成自动筛选操作后，再次单击“数据→排序和筛选→筛选”，将退出自动筛选状态，变量名旁边的筛选箭头也一起消失。

4.9.3　高级筛选

对于更为复杂的筛选，自动筛选可能无法完成筛选任务。例如，自动筛选可以对多个变量同时设置筛选条件，但是各变量筛选条件之间的关系只能是“与”的关系。又例如，自定义选项虽然功能较强，但是最多只能应用两个运算符。所以对于更为复杂的筛选，必须使用高级筛选实现。高级筛选（advanced filter）操作的关键是条件区域的设置。

应将条件区域（criteria range）放在整个数据清单的下方（以防止被筛选隐藏），条件区域的格式是这样的：第一行为变量名行，第二行及以下各行为相应的条件值，其中同一行条件间的关系为“与”的关系，不同行的条件间的关系是“或”的关系，可以采用的条件符号有“>、<、≥、≤”等（见图 4-18）。

以光盘文件“04-06 高级筛选.xlsx”中的数据为例，进行高级筛选的操作如下：

① 单击“数据→排序和筛选→高级”。

② 在弹出的“高级筛选”对话框（见图 4-19）中选取需要筛选的数据区域（列表区域）“$A$1:$J$12332”及条件区域“$L$1:$O$3”，单击“确定”，即可从 12331 条记录中找到满足条件的 8 条记录（见图 4-20）。

为了从高级筛选结果状态恢复到原先的显示，可单击“数据→排序和筛选→清除”。

L	M	N	O
性别	年龄	住院天数	住院总费用
1	>60	>=100	>10000
2	>55	>=100	>10000

图 4-18　条件区域

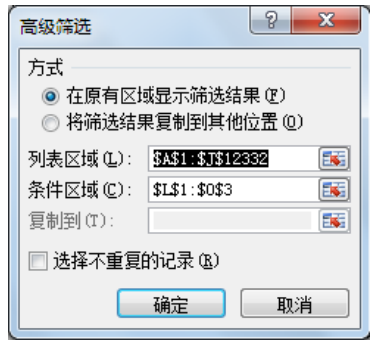


图 4-19　“高级筛选”对话框



	A	B	C	D	E	F	G	H	I	J
1	病案号	姓名	性别	年龄	疾病类别	入院日期	出院日期	住院天数	西药费	住院总费用
864	129978	胡国华	1	64	070.901+	1995/9/26	1996/4/4	191	49299.35	55896.75
2119	125369	宋东海	1	62		1995/10/25	1996/2/14	112	9367.60	13343.62
2311	132028	王福友	2	59	070.308	1995/12/4	1996/4/18	136	17672.71	23191.16
2400	134809	汤显超	1	65	070.308	1996/5/29	1996/9/19	113	24214.24	27894.29
8802	138276	何庭训	2	72	070.901	1995/9/12	1996/1/26	136	5083.60	12031.70
9076	116005	禾文东	1	78	070.106	1995/12/12	1996/4/18	128	40183.90	78764.64
10576	106463	张忠敏	1	65	571.402	1995/4/25	1995/11/2	191	44450.95	49945.23
10715	730728	郭健	1	80	414.001	1996/5/6	1996/9/27	144	16285.72	26654.43

疾病库 Sheet1

就绪　在 12331 条记录中找到 8 个　100%

图 4-20　高级筛选结果

4.9.4　选择不重复的记录

有时，一个观察个体有两个或多个值，但实际工作中只需要一个即可，即需要选择不重复的记录。Excel 通过高级筛选方式可简单地解决此问题。

光盘数据文件“04-07 不重复显示.xlsx”中的某一个观察个体重复了三次，只选择不重复记录的操作如下：

①单击“数据→排序和筛选→高级”，弹出“高级筛选”对话框（见图 4-19）。

②选取“列表区域”，以该区域为准来选择不重复记录。“列表区域”的选取可以是整个工作表数据区域，也可以是工作表数据区域的某一个变量或几个变量。

③单击“高级筛选”对话框左下角的“选择不重复的记录”复选框，单击“确定”按钮，即可从 24 个记录中筛选出 3 个记录。

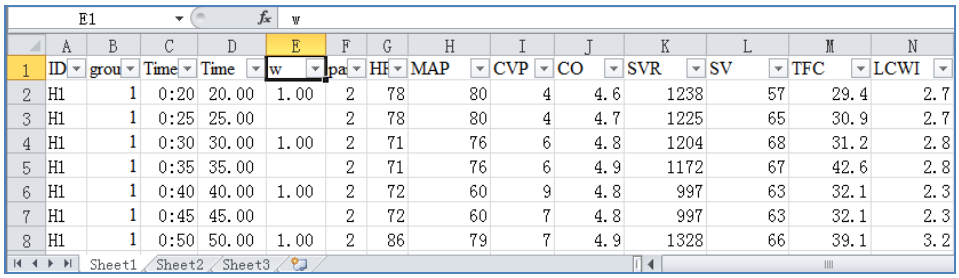
4.9.5　人工筛选

有时，采用上述自动或高级筛选方式不能满足筛选要求，可使用人工筛选方法。该方法的要点是：产生一个新的变量，该变量在需要挑选记录的单元格中为 1，其余为空白；如果间距相等，则可采用向下拖放填充的方式快速产生该变量；以此变量进行自动筛选。

例 4-5：对于光盘数据文件“04-08 人工筛选.xlsx”，为了将此数据清单一分为二，要求每间隔一个记录挑选一个，请给出挑选方案。

解：在实际数据分析中，类似于例 4-5 的情况较多，比如将一个数据清单按一定间隔取出记录、按事先规定比例从数据清单中随机抽取记录（参见“第 14 章随机化分组与样本含量估计”）等。就例 4-5 而言，需要新建一个变量（如 w），该变量在需要取出记录所在单元格为 1，其他单元格为空白（见图 4-21）。然后单击功能区中的“数据→排序和筛选→

筛选”；单击变量 w 右侧的下拉箭头按钮，在列表框中选定所需筛选的条件，即选定“1.00”（见图 4-22），此时即可达到筛选要求。



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	ID	grou	Time	Time	w	pa	HF	MAP	CVP	CO	SVR	SV	TFC	LCWI
2	H1	1	0:20	20.00	1.00	2	78	80	4	4.6	1238	57	29.4	2.7
3	H1	1	0:25	25.00		2	78	80	4	4.7	1225	65	30.9	2.7
4	H1	1	0:30	30.00	1.00	2	71	76	6	4.8	1204	68	31.2	2.8
5	H1	1	0:35	35.00		2	71	76	6	4.9	1172	67	42.6	2.8
6	H1	1	0:40	40.00	1.00	2	72	60	9	4.8	997	63	32.1	2.3
7	H1	1	0:45	45.00		2	72	60	7	4.8	997	63	32.1	2.3
8	H1	1	0:50	50.00	1.00	2	86	79	7	4.9	1328	66	39.1	3.2

图 4-21　在原数据清单中插入新变量 w

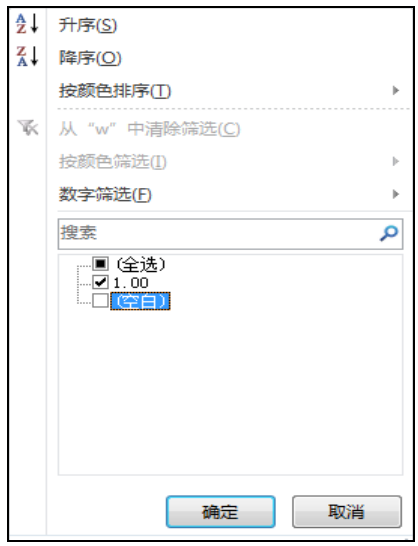


图 4-22　自动筛选新变量 *w*

对于图 4-12 的“合并计算”结果（见光盘文件“04-03 多个工作表的合并计算.xlsx”），如果单击按钮 2，此时可见：1983~2002 年 6 所医院三个指标的数值及其合并计算值；为了去除合并计算值，只保留 1983~2002 年 6 所医院三个指标的数值，可在第 E 列创建一个变量，使该变量对应于每年 6 所医院源数据的单元格数字为 1，合并值处为空白（注意：不能为“0”），如图 4-23 所示（即 E2:E7 区域值为 1，E8 单元格为空白）。选取单元格区域 E2:E8，向下拖放填充至单元格 E141。单击“数据→排序和筛选→筛选”，在单元格 E1 右侧单击下拉箭头，在下拉列表中选 1，即可筛选出所需资料。



	A	B	C	D	E
1	年份	出院人数	出院病人数	出院者占用总床日数	
2		4529	3921	60328	1
3		1123	876	18424	1
4		3307	3121	52628	1
5		11079	10218	193775	1
6		15463	14471	286976	1
7		31545	27895	460080	1
8	2002	67046	60502	1072211	
9		1732	1542	35379	1

图 4-23　人工筛选“合并计算”的多个工作表源数据

4.10　两个类似文件的比较

很多严格要求的科学研究，为了避免输入错误，通常需要输入双份数据，为了比较两份输入文件是否有所不同，在 Excel 中可以采用重排“窗口”。

1. 逐行检查

如果已经打开了光盘文件“04-01 妊娠分娩.xlsx”（注意先单击“取消冻结窗口”），然后再打开光盘文件“03-04 妊娠分娩.xlsx”，单击“视图→窗口→并排查看”后，出现如图 4-24 所示的界面。在两个上下界面中滑动滚动条可逐行比较每条记录。注意，处于当前界面的文件才出现水平或垂直滚动条，另一个非活动文件的滚动条消失。在图 4-24 中，“04-01 妊娠分娩.xlsx”为当前界面，如果鼠标移至“03-04 妊娠分娩.xlsx”，则“03-04 妊娠分娩.xlsx”变为当前界面。如果要停止“并排比较”，可再单击一次“并排查看”。



03-04 妊娠分娩 [兼容模式]

	A	B	C	D	E	F	G	H	I
1	住院号	年龄（y）	身高（cm）	体重（kg）	住院天数	职业	文化程度	分娩方式	妊娠结局
2	20080001	25	162	76.0	9	其他	中学	顺产	早产
3	20080002	32	153	60.0	7	其他	小学	剖宫产	足月
4	20080003	28	158	64.0	10	其他	中学	顺产	足月
5	20080004	29	162	68.0	8	工人	大学	剖宫产	足月
6	20080005	27	158	68.0	6	农民	小学	顺产	早产
7	20080006	39	158	66.5	8	工人	中学	剖宫产	早产
8	20080007	23	162	68.0	11	其他	小学	剖宫产	早产

妊娠分娩

04-01 妊娠分娩 [兼容模式]

	A	B	C	D	E	F	G	H	I
1	住院号	年龄（y）	身高（cm）	体重（kg）	住院天数	职业	文化程度	分娩方式	妊娠结局
2	20080001	25	162	76.0	9	其他	中学	顺产	早产
3	20080002	32	153	60.0	7	其他	小学	剖宫产	足月
4	20080003	28	158	64.0	10	其他	中学	顺产	足月
5	20080004	29	162	68.0	8	工人	大学	剖宫产	足月
6	20080005	27	158	68.0	6	农民	小学	顺产	早产
7	20080006	39	158	66.5	8	工人	中学	剖宫产	早产
8	20080007	23	162	68.0	11	其他	小学	剖宫产	早产

妊娠分娩

图 4-24　并排比较

2. 逐列检查

单击“视图→窗口→全部重排”，会弹出“重排窗口”对话框（如图 4-25），选择“垂直并排”，可以达到逐列比较的目的。

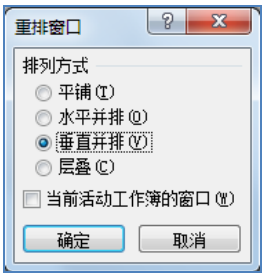


图 4-25 “重排窗口”对话框

第 5 章

分类汇总与透视分析

第 4 章的 4.8 节已经介绍了对多个 Excel 工作表数据进行合并计算的方法。本章将介绍对于同一个 Excel 工作表的数据，根据不同的字段，从不同角度对数据进行分析的方法，这一过程被形象地称为“透视分析”。透视分析是 Excel 最有特色的功能之一，看似复杂，但实际操作并不难。

5.1　分类汇总

分类汇总（subtotals）和合并计算（consolidate）类似，也可以完成求和、均数（即平均值）、连乘（即乘积）、最大值、最小值、非空白单元格的计数（即计数）、数字单元格的计数（即数值计数）、方差、总体方差、标准差（即标准偏差）、总体标准差（即总体标准偏差）11 项统计功能。

分类汇总是将数据清单中关键字相同的一些记录合并为一组，对于每组记录按列进行求和、计算均数等汇总。为了避免分类汇总计算的错误，在执行“分类汇总”命令前，应将数据清单中的记录按照某一关键字进行排序，然后才能根据此关键字进行汇总。

1. 简单分类汇总

例 5-1：某研究者将病人随机分成对照组、实验 1 组、实验 2 组三个组（分别记为 0、1、2），给每组不同的药物后，每相隔一定的时间测量一次病人的心率、血压、CVP、CO、SVR、SV、TFC、LCWI 等指标，获得的数据如图 5-1 所示（也见光盘文件“05-01 心率血压.xlsx”）。试采用分类汇总方法计算每组的心率、血压、CVP、CO、SVR、SV、TFC、LCWI 等指标的均数与标准差。

	A	B	C	D	E	F	G	H	I	J	K
1	编号	组别	时间	心率	血压	CVP	CO	SVR	SV	TFC	LCWI
2	B1	0	0:00	99	89	6.5	5.5	1210	55	34	4
3	B2	0	0:00	59	84	5	4.5	1401	68	34.3	2.8
4	B3	0	0:00	62	97	7	5.4	1498	80	44.6	4.8
5	H1	1	0:00	89	96	4	6.8	1063	77	44	4.9
6	H2	1	0:00	95	103	9.5	6.5	1139	62	35	5
7	H3	1	0:00	62	105	6	4.6	1858	54	39.9	4.7
8	G1	2	0:00	90	98	8	5	1301	52	38	3.3
9	G2	2	0:00	77	84	12	5.8	1318	57	39.4	4.7
10	G3	2	0:00	77	71	8	6.1	1119	85	29	5.1
11	B1	0	0:05	105	93	6.5	5.7	1238	59	34.3	4.4
12	B2	0	0:05	56	81	4.5	4.4	1417	79	34.1	2.8
13	B3	0	0:05	58	115	7	5.4	2053	87	45	6.5
14	H1	1	0:05	92	85	4	6.8	1066	78	44	4.9
15	H2	1	0:05	105	67	8.5	6.1	1205	59	35.4	4.7
16	H3	1	0:05	80	89	7	3.7	2231	46	40.4	3.6
17	G1	2	0:05	100	87	8	5.1	1271	50	31.6	3.4
18	G2	2	0:05	74	72	12	5.2	1498	72	38.9	4.2

图 5-1　某药物代谢动力学研究数据

解：分类汇总计算均数可按如下步骤进行。

① 对需要分类汇总的关键字字段进行排序。本例应按“组别”进行排序。单击图 5-1 中的“组别”，即选中单元格 B1，单击“开始→编辑→排序和筛选”中的 升序(S) 按钮（亦可通过单击“数据→排序和筛选→升序”完成），即对整个数据清单按“组别”升序排序。

如果不排序，汇总结果将可能会零乱分散，达不到预期目的。

② 选定数据清单中的任一单元格。单击“数据→分级显示→分类汇总”，弹出“分类汇总”对话框，如图 5-2 所示。

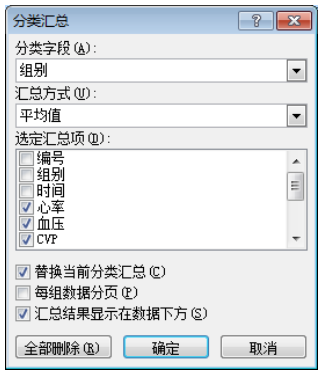


图 5-2 “分类汇总”对话框及部分数据清单数据

③ 单击“分类字段”框的向下箭头，在弹出的下拉列表中选择所需字段作为分类汇总的依据，这里指定“组别”（注意：一定要排序）为分类字段。

④ 单击“汇总方式”框的向下箭头，在弹出的下拉列表中有求和、计数等 11 个选项。为了获得均数，这里指定“平均值”（如果要获得标准差，应指定“标准偏差”）。

⑤ 在“选定汇总项”列表框中，列出了数据清单中的所有字段（变量）名称，将所有需要汇总的字段前面复选框选中。这里指定心率、血压、CVP、CO、SVR、SV、TFC、LCWI 等字段。

⑥ 根据需要，指定汇总结果显示的位置，选定相应的复选框。其中，“替换当前分类汇总”表示按本次分类汇总要求进行汇总；“每组数据分页”表示每一类按分页显示，当数据很多时，这样处理有助于查看；“汇总结果显示在数据下方”表示将分类汇总数放在本类的最后一行，系统默认的方式是将分类汇总数放在本类的第 1 行。本例选中“替换当前分类汇总”和“汇总结果显示在数据下方”（见图 5-2）。

⑦ 单击“确定”按钮，获得如图 5-3 所示的结果。

	A	B	C	D	E	F	G	H
1	编号	组别	时间	心率	血压	CVP	CO	SVR
2	B1	0	0:00	99	89	6.5	5.5	1210
3	B2	0	0:00	59	84	5	4.5	1401
4	B3	0	0:00	62	97	7	5.4	1498
5	B1	0	0:05	105	93	6.5	5.7	1238
6	B2	0	0:05	56	81	4.5	4.4	1417
7	B3	0	0:05	58	115	7	5.4	2053
8	B1	0	0:10	98	74	6.5	4.8	1459

Sheet1

图 5-3 分类汇总结果

在图 5-3 的左上角汇总分为 1、2、3 级，分别表示总的汇总、分组汇总、明细数据。为了得到按组别分组后的汇总结果，应单击汇总 2 级，此时显示结果如图 5-4 所示。

	A	B	C	D	E	F	G	H	I	J	K
1	编号	组别	时间	心率	血压	CVP	CO	SVR	SV	TFC	LCWI
57		0 平均值		81.2182	79.3455	6.5455	4.802	1265.6	62.05	36.071	3.130909
130		1 平均值		87.1667	80.1389	7.3542	5.874	1078.7	66.21	38.772	3.826761
204		2 平均值		78.9863	72.1507	13.234	5.282	1060.2	67.78	36.974	3.20411
205		总计平均值		82.545	77.005	9.278	5.363	1123.3	65.64	37.373	3.40603

图 5-4　分组汇总的结果

此时，Excel 将为每组记录插入一个分类汇总行，在该行相应单元格利用分类汇总函数“SUBTOTAL(function_num,ref1,ref2,...)”，按要求计算分类汇总值。其中“function_num”的取值为 1 到 11 之间的数字，分别表示 11 个函数 AVERAGE、COUNT、COUNTA、MAX、MIN、PRODUCT、STDEV、STDEVP、SUM、VAR、VARP。“ref1,ref2,...”为引用数值或引用单元格。如本例，求对照组心率均数的公式为“=SUBTOTAL(1,D2:D56)”，单击 D57 单元格（显示均数为：81.2182）可看到此公式。

在图 5-4 左侧的符号 ⊞ 为显示明细数据按钮，而 ⊟ 为隐藏明细数据按钮。要显示某一个分组数据的明细数据，可单击相应汇总数据所对应的显示明细数据按钮 ⊞，这时相应的按钮变为隐藏明细数据按钮 ⊟。为了显示对照组（即 0 组）的明细数据，在第 57 行左侧单击明细数据按钮 ⊞，可获得如图 5-5 所示的结果。单击级别 3 按钮 3 可显示所有组的明细数据。

D57　fx　=SUBTOTAL(1,D2:D56)

	A	B	C	D	E	F	G	H	I	J	K
53	B1	0	3:35	91	83	7	4.1	1581	46	27	3
54	B2	0	3:35	84	75	7	5.2	1041	64	37.2	2.8
55	B2	0	3:50	85	83	7	6.1	969	71	37.6	3.6
56	B2	0	4:05	82	90	7	5.2	1293	63	38.9	3.5
57		0 平均值		81.2182	79.3455	6.5455	4.802	1265.6	62.05	36.071	3.130909
130		1 平均值		87.1667	80.1389	7.3542	5.874	1078.7	66.21	38.772	3.826761
204		2 平均值		78.9863	72.1507	13.234	5.282	1060.2	67.78	36.974	3.20411
205		总计平均值		82.545	77.005	9.278	5.363	1123.3	65.64	37.373	3.40603

图 5-5　查看对照组明细数据

要将分类汇总的结果单独存放在一张工作表中，可将汇总的结果复制后，利用编辑菜单中的“选择性粘贴”命令粘贴到新工作表中。在“选择性粘贴”对话框的“粘贴”选项中，应选择“数值”选项。为了删除明细数据，可利用筛选操作进行筛选，或直接删除。

为了获得该数据的标准差，可仿照上述步骤进行。只要将上述第 4 步中的“汇总方式”指定为“标准偏差”即可。

2. 删除分类汇总

如果要删除分类汇总，可单击“数据→分级显示→分类汇总”，然后在“分类汇总”对话框中单击“全部删除”。此时，数据清单恢复成原来的初始状态。

3. 多级分类汇总

如果需要，还可以在简单分类汇总的基础上再创建二级乃至多级分类汇总。多级分类汇总的关键是正确的分类汇总次序，即先按主要关键字建立第 1 级分类汇总，再按次要关键字建立第 2 级分类汇总，依次类推。在分类汇总前必须按主要关键字、次要关键字等排序。

例 5-2：对于例 5-1 资料，按“组别”和“时间”两个字段求心率、血压、CVP、CO、SVR、SV、TFC、LCWI 等指标的最大值。

解：按以下步骤操作：

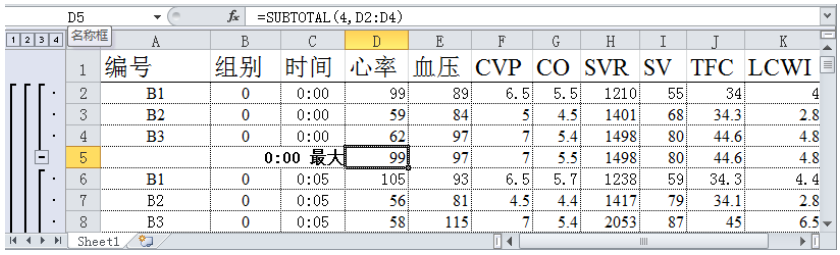
①删除分类汇总。

②在菜单中单击功能区内“开始→编辑→排序和筛选→自定义排序”，主要关键字选“组别”，单击“添加条件”按钮，次要关键字选“时间”，均为升序，单击“确定”按钮。

③按主要关键字“组别”建立第 1 级分类汇总，方法同上。注意：在“汇总方式”框中指定“最大值”。

④按次要关键字“时间”建立第 2 级分类汇总，方法同上。只是必须清除“替换当前分类汇总”复选框中的选定标志。

这样，按“组别”和“时间”分组进行二级分类汇总的结果如图 5-6 所示。



	A	B	C	D	E	F	G	H	I	J	K
1	编号	组别	时间	心率	血压	CVP	CO	SVR	SV	TFC	LCWI
2	B1	0	0:00	99	89	6.5	5.5	1210	55	34	4
3	B2	0	0:00	59	84	5	4.5	1401	68	34.3	2.8
4	B3	0	0:00	62	97	7	5.4	1498	80	44.6	4.8
5			0:00 最大	99	97	7	5.5	1498	80	44.6	4.8
6	B1	0	0:05	105	93	6.5	5.7	1238	59	34.3	4.4
7	B2	0	0:05	56	81	4.5	4.4	1417	79	34.1	2.8
8	B3	0	0:05	58	115	7	5.4	2053	87	45	6.5

图 5-6　二级分类汇总结果

图 5-6 中数据是级别按钮 4 的汇总数据；如果单击级别 3 按钮 3，可显示按“组别”和“时间”交叉的汇总数据；如果单击级别 2 按钮 2，则仅显示按“组别”的汇总结果。

5.2　数据透视表

数据透视表（pivottable）是“数据”菜单中“数据透视表与数据透视图”的其中一部

分，数据透视图将在下一节介绍。“数据透视表与数据透视图”有机地综合了前面介绍的数据排序、筛选、分类汇总等数据分析的优点，可方便地调整分类汇总的方式（参见 5.1 节），灵活地以多种不同方式展示数据的特征。计算均数标准差（分类汇总）、建立列联表、计算百分比、建立新的数据子集、绘制统计图形等均可利用这一工具完成。因此，该工具是最常用、功能最全的 Excel 数据分析工具之一。掌握这一功能，可以帮助用户解决许多数据分析问题。

例 5-3：某医院妇产科医生收集了 1402 例产妇分娩资料（前面两章已用过此数据，见光盘文件“03-04 妊娠分娩.xlsx”或“04-01 妊娠分娩.xlsx”），住院号是识别每个产妇的编号；年龄、身高、体重、住院天数 4 个变量为定量数据（数字数据），其度量单位分别为岁（y）、厘米（cm）、公斤（kg）、天（d）；职业（工人、农民、商业服务、知识分子、管理人员、其他）、文化程度（文盲、小学、中学、大学）、分娩方式（顺产、先兆早产、助产、剖宫产）、妊娠结局（足月、早产、过期产、死产）4 个变量为定性数据（文本数据）。请利用这一数据清单进行数据透视分析。

5.2.1　创建数据透视表

创建数据透视表的具体步骤如下：

① 打开光盘文件“03-04 妊娠分娩.xlsx”，在该数据工作表中选取任一单元格，以便 Excel 自动确定用户想用的数据范围，单击“插入→表→数据透视表▼→数据透视表”，弹出“创建数据透视表”对话框（见图 5-7）。

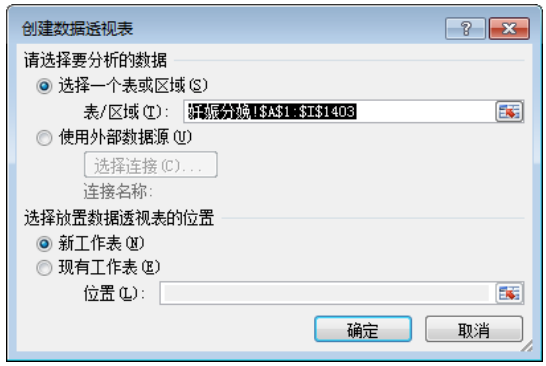


图 5-7　创建数据透视表

若要创建数据透视图，只需单击“数据透视表▼→数据透视表”。

② 在“请选择要分析的数据”中，确定 Excel 已正确识别数据的单元格区域 $A$1:$I$1403（一般都能自动识别。如果没有正确选定区域，应使用鼠标单击框右侧箭头，

进行单元格区域的选取）；亦可使用外部数据源。

③在“选择放置数据透视表的位置”中，指定数据透视表显示的位置，本例按默认选定“新工作表”。

④单击“确定”按钮，Excel 自动创建一张新的工作表（Sheet1）（见图 5-8）。数据透视表工具分为“选项”、“设计”两个功能区。“选项”功能区包括了“排序和筛选”、“数据”、“操作”、“计算”、“工具”、“显示”等命令组；而“设计”功能区包括了“布局”、“透视表样式选项”、“透视表样式”等命令组。Excel 工作表左侧为生成数据透视表结果的区域（Excel2003 版本在此处也包括了透视表的设计）；右侧上方为“数据透视表字段列表”，右侧下方为透视表设计区，即通过拖动字段名到“数值”、“行标签”、“列标签”、“报表筛选”区域来设计生成数据透视表的内容。

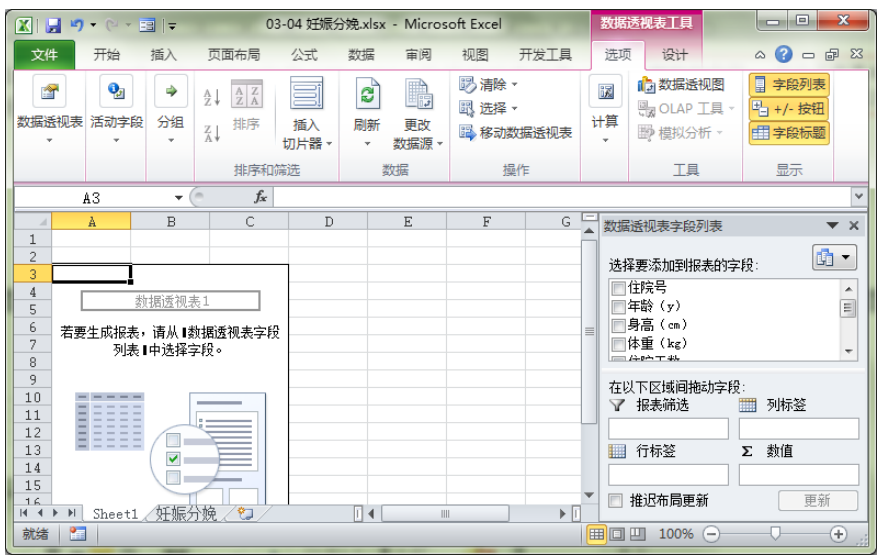


图 5-8　数据透视表调用时的初始形式

温馨提示：Excel 2003 调用数据透视表是单击“数据→数据透视表和数据透视图”，其他过程与 2010 版基本相同。在 Excel 2010 中，读者可通过在键盘上依顺序按“Alt”、“D”、“P”三个键调用 2003 版中的“数据透视表和数据透视图向导”。

5.2.2　数据透视表的结构及其工具栏

1. 4 个区域

在设计数据透视表时，需要明确 4 个区域（areas）。这 4 个区域 Excel 2003 以前版本位

于左侧，而 Excel 2010 位于右下侧。

（1）“**数值（value）**”区，在图 5-8 中在最右下角的位置。为了计数或求频数，一般将个体编号（如住院号）变量拖入此处；为了计算定量数据描述性统计量（如平均值、标准差等），应将定量数据（如年龄、身高、体重、住院天数）拖入此处。

（2）“**行标签（row labels）**”区，该区域一般应拖入分类变量，可拖入多个分类变量，多个变量依次从上到下排列。

（3）“**列标签（row labels）**”区，该区域一般也应拖入分类变量，也可拖入多个分类变量，多个变量依次从上到下排列。

（4）“**报表筛选（report filter）**”区，该区域位于这 4 个区域的左上角。如果将分类变量拖放在此处，选择这一变量的某一类别，则产生一个数据透视表，选择这一变量的另一类别，则覆盖前次产生的表格，产生一个新的数据透视表；选取“（全部）”时，则忽略该变量的作用。如果需要同时选择这一变量的多个类别，可在选项列表的下方“选择多项”前打对钩。这一区域的作用相当于统计上所说的分层，即按拖入的变量分层后再看各层的数据分析结果。

2. 数据透视表工具

Excel 2010 的“数据透视表工具”，含有“选项”（见图 5-8）和“设计”（见图 5-9）两个选项卡。如果没有看到“数据透视表工具”，只要单击左侧的生成数据透视表的任何区域便可出现该工具（“数据透视表字段列表”以及上面介绍的 4 个区域也会同时出现）。



图 5-9　数据透视表工具中的“设计”选项卡

3. 编辑与组织数据透视表

创建了数据透视表雏形（见图 5-8）后，应根据分析要求，设置数据透视表的版式。该步骤也是创建数据透视表最关键的一步。

如果要将某个变量添加到行标签区、列标签区、报表筛选区或数值区时，则将相应变量拖放到这 4 个区即可。如果要从数据透视表中删除某个变量，则将该变量拖放回数据透视表字段列表中即可。在用鼠标拖动某个变量时，鼠标指针会随着鼠标指针所处的位置而变化，提示用户这是放开鼠标按键时的操作结果。当鼠标位于行标签区、列标签区、报表筛选区、数值区、以及将变量拖回删除时，各种鼠标指针的形状依次为图 5-10 所示。

如果没有看到“数据透视表字段列表”，只要在左侧单击生成数据透视表的任何区域即

可。也可单击“选项→显示/隐藏→字段列表”。

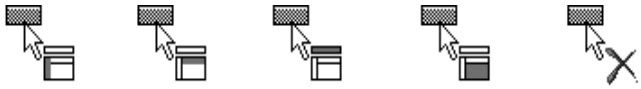


图 5-10　各种鼠标指针的形状

5.2.3　由数据透视表获得均数标准差

例 5-4：对于例 5-3 资料，按不同妊娠结局分组，分别计算年龄、身高、体重、住院天数 4 个定量数据的各组均数标准差。

解：采用数据透视表可按下列操作步骤计算均数标准差。

① 在存放例 5-3 资料的工作表（即“03-04 妊娠分娩.xlsx”文件）中，选取任一单元格后，单击“插入→表格→数据透视表”，再单击“确定”按钮，获得数据透视表雏形（见图 5-8）。

② 从数据透视表字段列表中，将“妊娠结局”拖放至行标签区，将年龄、身高、体重、住院天数分别拖放至数值区，此时获得图 5-11 的界面。

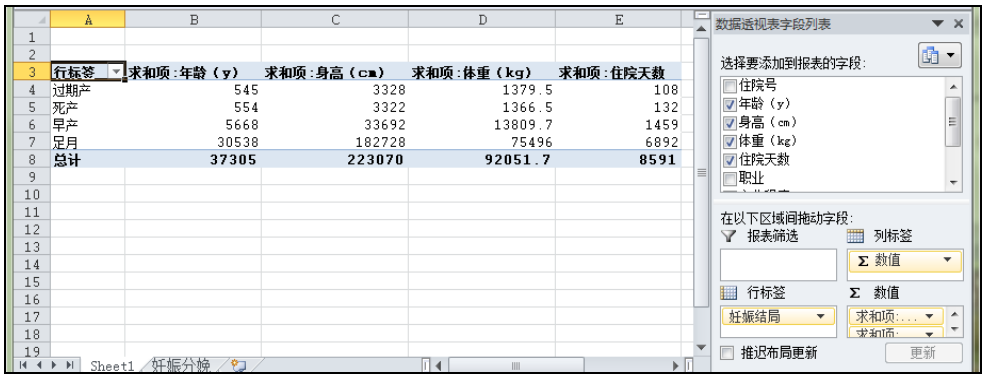


图 5-11　选取行字段与数据项字段后的界面

③ 先选取单元格 B3 的“求和项:年龄（y）”，然后快速双击此单元格；或在右下方数值区单击“求和项:年龄（y）→值字段设置”。以上两种情况下即可弹出“值字段设置”对话框（见图 5-12），在“值汇总方式”下拉列表框中选择“平均值”。单击正下方的“数字格式”，弹出“设置单元格格式”对话框，此对话框与第 3 章介绍的“数字格式”相同（参见 3.1 节），在“分类”列表框中选择“数值”，小数位数设置为 1 位，符号设置为“-1234.0”形式。双击“确定”按钮，此时便可获得按妊娠结局分组后的年龄平均值。要想获得身高、体重、住院天数的平均值，可仿照年龄的平均值计算来进行。完成这些计算后的数据透视

表结果见图 5-13 所示。

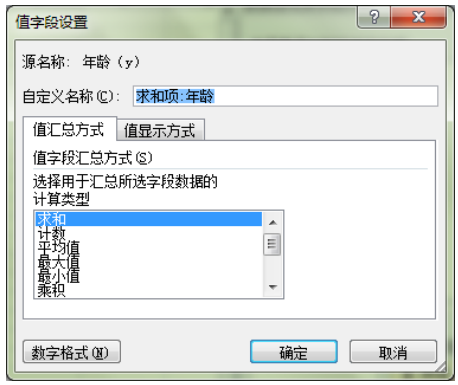


图 5-12　“值字段设置”对话框

行标签	平均值项:年龄	平均值项:身高	平均值项:体重	平均值项:住院
过期产	26.0	158.5	65.7	5.1
死产	26.4	158.2	65.1	6.3
早产	26.7	158.9	65.1	6.9
足月	26.6	159.2	65.8	6.0
总计	**26.6**	**159.1**	**65.7**	**6.1**

图 5-13　按妊娠结局分组后的年龄等 4 个变量的平均值

为了计算按妊娠结局分组后的年龄等 4 个变量的标准差，可单击“数据透视表工具→选项→计算→按值汇总→其他选项”，在弹出的“值字段设置”（见图 5-12），修改“平均值”为“标准偏差”，单击“确定”按钮即可。当然，采用前面求解年龄平均值的方法，同样可以获得这 4 个变量的标准差（读者不妨一试）。结果见图 5-14 所示。

行标签	标准偏差项:年龄	标准偏差项:身高	标准偏差项:体重	标准偏差项:住院
过期产	3.9	4.5	6.2	1.8
死产	3.1	2.6	5.8	3.8
早产	3.7	3.6	5.8	5.0
足月	3.2	4.1	6.1	2.8
总计	**3.3**	**4.0**	**6.1**	**3.2**

Sheet1　妊娠分娩

图 5-14　按妊娠结局分组后的年龄等 4 个变量的标准差

“值汇总方式”列出了求和、均数（即平均值）、连乘（即乘积）、最大值、最小值、非空白单元格的计数（即计数）、数字单元格的计数（即数值计数）、方差、总体方差、标准差（即标准偏差）、总体标准差（即总体标准偏差）11 项统计功能，与前面介绍的“分类

汇总”相同。

单击数据透视表区域中“行标签”右侧下拉箭头，将弹出一个“妊娠结局”分类类别列表框（见图 5-15），从中可选取需要显示的类别或统计指标。此外也可对分类变量进行排序。

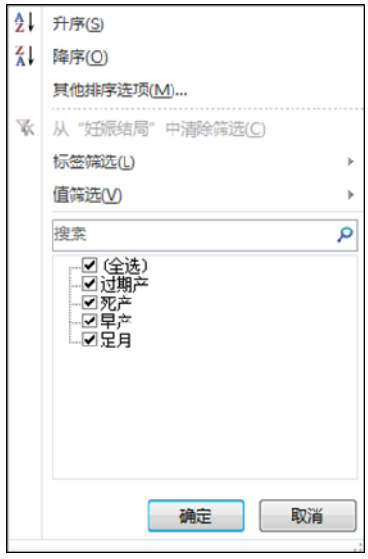


图 5-15 “妊娠结局”类别

5.2.4 产生列联表

1. 一维列联表

例 5-5：建立定性变量“职业”的频数与相对频数。

解：获得按“职业”分组的频数及其相对频数的步骤如下：

① 在存放例 5-3 资料的工作表（即“03-04 妊娠分娩.xlsx”文件）中，选取任一单元格后，单击“插入→表格→数据透视表”，然后单击“确定”按钮，获得数据透视表雏形（见图 5-8）。先从字段列表中将变量“职业”拖入“行标签”区；然后将变量“住院号”拖入字段区域右下角处“数值”区，因为每个个体有一个“住院号”，所以可利用它进行计数。此时，数据透视表及其字段列表如图 5-16 所示。

② 按“职业”分组后清点各类别的例数。本例的目的不是对“住院号”求和，而是计数，因此需要修改。单击“数值”区的“求和项:住院号→值字段设置”，在弹出的“汇总方式”下拉列表框中选择“计数”，单击“确定”按钮，获得的结果见表 5-1 的第 2 列。

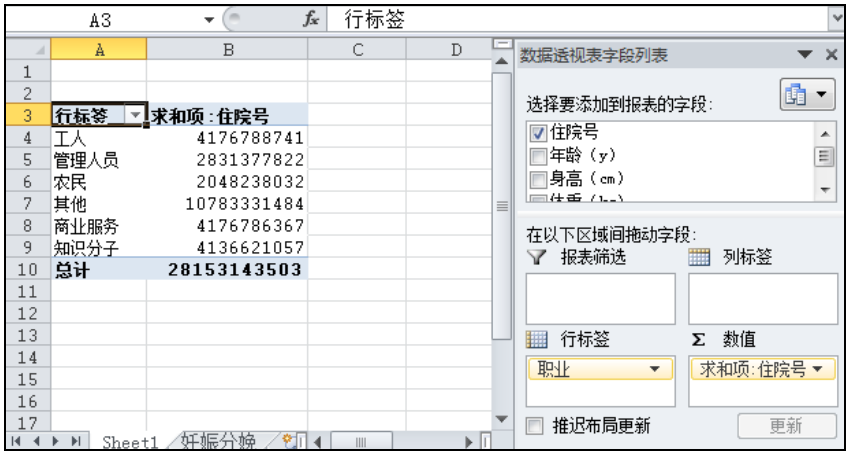


图 5-16　改变的数据透视表及其字段列表

表 5-1　“职业”的频数表

职业 （1）	频数 （2）	相对频数（%） （3）
工人	208	14.84
管理人员	141	10.06
农民	102	7.28
其他	537	38.30
商业服务	208	14.84
知识分子	206	14.69
合计	1402	100.00

③ 为了获得表 5-1 的第 3 列，可在如图 5-17 所示的“值字段设置”对话框中单击“值显示方式”。在“值显示方式”下方右侧单击下箭头选择“列汇总的百分比”，单击“确定”按钮，获得如图 5-17 左侧数据透视表区域所示的结果。

“值显示方式”列表框中的选项及其功能如表 5-2 所示。其中“基本字段”是指数据清单中的字段（变量），如身高（cm）、体重（kg）、住院天数、职业等（见图 5-17）。“基本项”包括所选定“基本字段”的取值及“上一个”、“下一个”。“上一个”表示与当前值之前的数值作比较；“下一个”表示与当前值之后的数值作比较。

2. 二维列联表

二维列联表可反映两个分类变量（定性变量）之间的关系，后面将要提到 χ^2 检验，这种检验的四格表、行×列表（也称 $R \times C$ 表）就是二维列联表。列联表（crosstab）是如何

根据原始数据产生的呢？见下面的例 5-6。

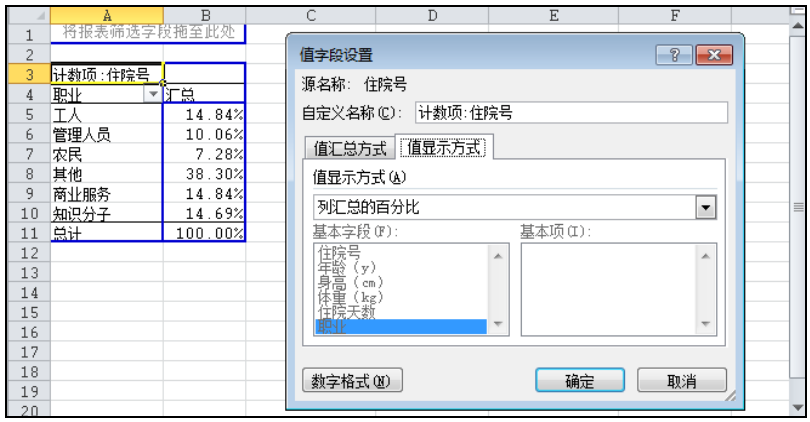


图 5-17　列联表结果及“值字段设置”对话框

表 5-2　数据显示方式及其功能

选　项	功　能
无计算	一般方式显示
差异	显示数值区所有数据与指定的基本字段和基本项数值之差
百分比	显示数值区所有数据与指定的基本字段和基本项数值的百分比
差异百分比	与“差异”相同，但差异以百分数形式显示
按某一字段汇总	为连续型字段项显示连续的总计，必须为要显示的项选择字段
行汇总的百分比	列联表的每行中各个数据与该行总计的百分比（行百分数之和为 100%）
列汇总的百分比	列联表的每列中各个数据与该列总计的百分比（列百分数之和为 100%）
全部汇总的百分比	列联表的各格子数据与总例数的百分比（各格子百分数之和为 100%）
指数	公式为：[单元格的值×总计]/[行总计×列总计]

例 5-6:根据例 5-3 资料（即“03-04 妊娠分娩.xlsx”文件），建立变量“文化程度”与“分娩方式”的二维列联表。

解：获得二维列联表的具体操作步骤如下：

① 在存放例 5-3 资料的工作表中，选取任一单元格后，单击“插入→表格→数据透视表”，然后单击“确定”按钮，获得数据透视表雏形（见图 5-8）。

② 先将变量“文化程度”拖入“行标签”区，将变量“分娩方式”拖入“列标签”区；然后将变量“住院号”拖入“数值”区。

③ 单击字段区域中“数值”区的“求和项:住院号→值字段设置”，在弹出的“汇总方式”下拉列表框中选择“计数”，单击“确定”按钮，即可获得变量“文化程度”与“分娩

方式”的二维列联表（见图 5-18）。

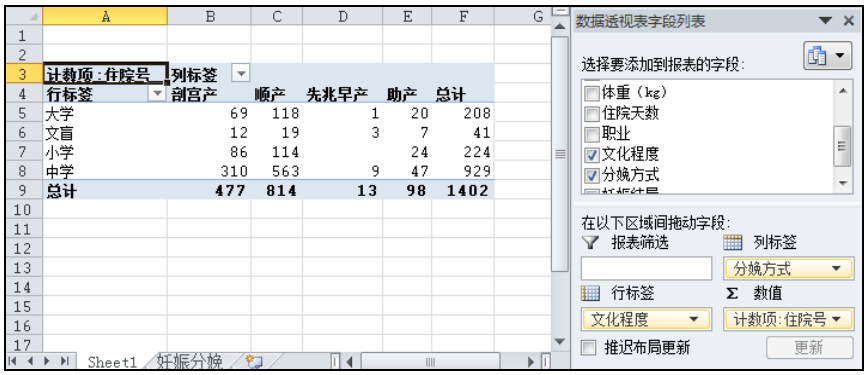


图 5-18　文化程度与分娩方式的二维列联表

④为了将数据透视表获得的列联表修改成常用的统计表形式，可通过复制、选择性粘贴（数值），并对部分分类类别的顺序关系略加调整，可得到所需的二维列联表（见表 5-3）。

表 5-3　文化程度与分娩方式之间的关系

文化程度	分娩方式				
	顺　产	先兆早产	助　产	剖 宫 产	合　计
文盲	19	3	7	12	41
小学	114	0	24	86	224
中学	563	9	47	310	929
大学	118	1	20	69	208
合计	814	13	98	477	1402

“字段设置”界面与“值字段设置”界面显示有所不同。如单击字段区域中“列标签”区的“分娩方式→字段设置”，会弹出图 5-19 所示的“分娩方式”字段设置界面，比较图 5-19 与图 5-12（数据项“年龄”）可知，“行标签”或“列标签”属于分类变量，通常需要对其按一定顺序排列；而“数值”通常显示为各种数字，通常需要对数字格式或显示方式进行定义。

在以前版本中的行或列字段（标签）有一个“高级”按钮。单击“高级”按钮，可设置行或列字段的排列顺序，默认为手动排序，即可以拖动行或列字段的分组类别方式，重新排列分组类别的顺序。在 Excel 2010 中，此功能可以通过右击行或列标签中的各类别名称，在菜单中选择“排序”或“移动”来实现，弹出的界面如图 5-20 所示。

为了使图 5-18 中的“文化程度”和“分娩方式”的排列与表 5-3 相同，可通过“排序”

或“移动”（见图 5-20）来实现。

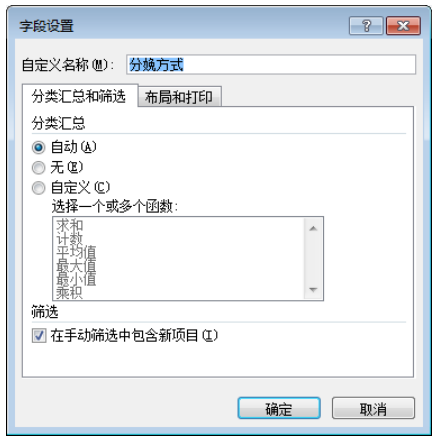


图 5-19　列字段“分娩方式”的字段设置界面

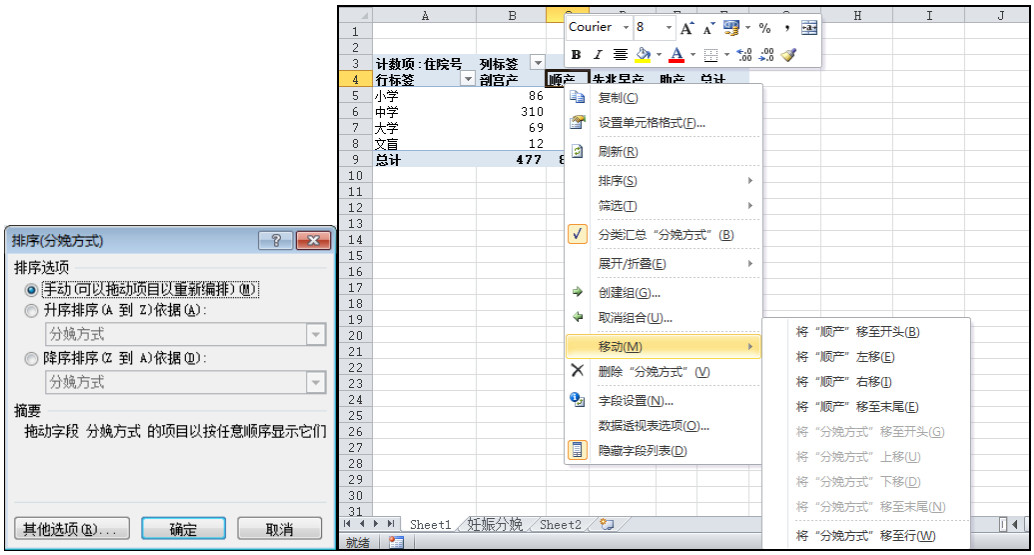


图 5-20　“排序”与“移动”界面

⑤为了比较不同文化程度的各分娩方式的构成情况，可在“值字段设置”对话框中单击“值显示方式”，并在“值显示方式”下单击倒三角箭头，选择“行汇总的百分比”，单击“确定”按钮即可。生成的结果见图 5-21。

结果分析：从图 5-21 中结果可见，在该医院分娩的妇女中，分娩方式以顺产为主（58.06%），其次是剖宫产，剖宫产在所有住院分娩妇女中占三分之一（34.02%）。文盲文

化程度的妇女中助产、先兆早产构成相对较多，顺产、剖宫产构成相对较少。

	A	B	C	D	E	F
1						
2						
3	计数项:住院号	列标签				
4	行标签	顺产	先兆早产	助产	剖宫产	总计
5	文盲	46.34%	7.32%	17.07%	29.27%	100.00%
6	小学	50.89%	0.00%	10.71%	38.39%	100.00%
7	中学	60.60%	0.97%	5.06%	33.37%	100.00%
8	大学	56.73%	0.48%	9.62%	33.17%	100.00%
9	总计	58.06%	0.93%	6.99%	34.02%	100.00%

图 5-21　不同文化程度的分娩方式构成比

3. 多维列联表

按同样方式也可建立多维列联表。

例 5-7：建立不同职业、不同文化程度与分娩方式的列联表。

解：获得满足上述要求的多维列联表可在上例的基础上进行。

在例 5-6 的基础上，先将“文化程度”拖回至字段列表之中，将其从“行标签”中删除。然后再将变量“职业”、“文化程度”依次拖入“行标签”区。变量“分娩方式”在“列标签”区保持不变（注意：这里已对“文化程度”与“分娩方式”进行了排序），变量“住院号”也已在“数值”区。单击字段区域中“数值”区的“计数项:住院号”，在弹出的列表中选取“值字段设置”，在“值显示方式”中将“行汇总的百分比”修改为“无计算”，单击“确定”按钮，即可获得不同职业、不同文化程度与分娩方式的三维列联表，如图 5-22 所示。

	A	B	C	D	E	F
3	计数项:住院号	列标签				
4	行标签	顺产	先兆早产	助产	剖宫产	总计
5	工人	115		10	83	208
6	文盲	3		2	4	9
7	小学	19		5	22	46
8	中学	84		3	51	138
9	大学	9			6	15
10	管理人员	83	2	14	42	141
11	小学	7			1	8
12	中学	39	2	5	29	75
13	大学	37		9	12	58
14	农民	56		4	42	102
15	文盲	5		2	4	11
16	小学	7		1	9	17
17	中学	43		1	26	70
18	大学	1			3	4
19	其他	322	1	41	173	537
20	文盲	7		2	1	10
21	小学	61		13	42	116
22	中学	245	1	25	121	392
23	大学	9		1	9	19
24	商业服务	124	7	16	61	208
25	文盲	4	3	1	3	11
26	小学	19		5	10	34
27	中学	97	4	9	47	157
28	大学	4		1	1	6
29	知识分子	114	3	13	76	206
30	小学	1			2	3
31	中学	55	2	4	36	97
32	大学	58	1	9	38	106
33	总计	814	13	98	477	1402

Sheet1 妊娠分娩

图 5-22　三维列联表

5.2.5 创建子数据清单

数据透视表可以创建汇总结果的明细数据清单，这一功能类似于前面介绍的筛选与查询。

从图 5-22 可以看出，职业为“工人”、文化程度为“中学”、分娩方式为“剖宫产”对应的三维列联表单元格为 E8，显示的“计数”为 51。该数值的含义是：在 1402 名孕产妇中，满足职业为“工人”、文化程度为“中学”、分娩方式为“剖宫产”条件者有 51 人。

双击单元格 E8，或单击鼠标右键选择“显示详细信息”，可获得这 51 个记录的明细数据，Excel 将这些数据自动存放在一个新的工作表中。图 5-23 显示了新产生的 51 个记录的前 10 个记录。

	A	B	C	D	E	F	G	H	I
1	住院号	年龄（y	身高（cm	体重（kg	住院天数	职业	文化程度	分娩方式	妊娠结局
2	20081399	26	159	65	6	工人	中学	剖宫产	足月
3	20081248	25	162	69	12	工人	中学	剖宫产	足月
4	20081245	26	163	66	9	工人	中学	剖宫产	足月
5	20081238	33	151	55	5	工人	中学	剖宫产	足月
6	20081221	25	160	68	7	工人	中学	剖宫产	足月
7	20080006	39	158	66.5	8	工人	中学	剖宫产	早产
8	20081220	29	156	63.5	8	工人	中学	剖宫产	足月
9	20081215	29	155	69	5	工人	中学	剖宫产	足月
10	20081168	26	157	58	8	工人	中学	剖宫产	足月
11	20081151	32	156	65	6	工人	中学	剖宫产	足月

图 5-23 51 个记录的明细数据的部分记录

双击感兴趣的其他汇总单元格数据（或单击鼠标右键选择“显示详细信息”），可继续产生满足条件的子数据清单，并自动存放到另一个新的工作表中。这一功能为 Excel 各层次用户的筛选或查询带来了很大方便。

为了清除第一行变量名后面的下拉箭头，单击“开始→编辑→排序和筛选→筛选”即可。

5.2.6 设置报告格式与更新数据

1. 设置数据透视表格式

与其他工作表一样，数据透视表也可以设置格式，达到美化修饰数据透视表的目的。

Excel 提供了多种“套用表格格式”（浅色、中等深浅、深色等）与“自动套用格式”（这个是 Excel 2003 的传统格式，Excel 2010 中一般不显示，需要单击“文件→选项→自定义功能区”设定）。如对图 5-22 中的表格选择“套用表格格式”中“中等深浅”的第 1 个格式，得到图 5-24。

其次，可通过单击“数据透视工具”的“选项”选项卡中的“显示”命令组，可对“+/-按钮”、“字段标题”进行隐含或显示修改。

	A	B	C	D	E	F	G
4	行标签	顺产	先兆早产	助产	剖宫产	总计	
5	工人	115		10	83	208	
6	文盲	3		2	4	9	
7	小学	19		5	22	46	
8	中学	84		3	51	138	
9	大学	9			6	15	
10	管理人员	83	2	14	42	141	
11	小学	7			1	8	
12	中学	39	2	5	29	75	
13	大学	37		9	12	58	
14	农民	56		4	42	102	
15	文盲	5		2	4	11	
16	小学	7		1	9	17	
17	中学	43		1	26	70	
18	大学	1			3	4	
19	其他	322	1	41	173	537	
20	文盲	7		2	1	10	
21	小学	61		13	42	116	
22	中学	245	1	25	121	392	
23	大学	9		1	9	19	
24	商业服务	124	7	16	61	208	
25	文盲	4	3	1	3	11	
26	小学	19		5	10	34	
27	中学	97	4	9	47	157	
28	大学	4		1	1	6	
29	知识分子	114	3	13	76	206	
30	小学	1			2	3	
31	中学	55	2	4	36	97	
32	大学	58	1	9	38	106	
33	总计	814	13	98	477	1402	

图 5-24　对图 5-22 套用“中等深浅色 1”格式后的数据透视表

2. 更新数据透视表

在创建了新的数据透视表后，如果在源数据清单中更改了某个数据，基于此数据清单的数据透视表并不会自动随之改变，需要更新数据。

在数据透视表中任意选择一个单元格，然后单击“数据透视表工具→选项→数据→刷新▼→全部刷新”（也可任意选择一单元格，单击鼠标右键后，单击“刷新”按钮）。这时，数据透视表中的数据将根据修改的源数据自动更新。

数据透视表与一般工作表相比，有如下两个特点：

①“透视”性。数据透视表可以根据用户的需要，对数据透视表的报表筛选字段、行标签字段、列标签字段、数值字段进行不同的设置，从多角度分析数据。此外，用户还可以改变汇总方式及显示方式，从而为用户分析数据提供了极大的方便。

②“只读”性。数据透视表可以像一般工作表一样进行修饰或制作图表，但有时不能达到“即改即可见”的效果。即在源数据清单中更改了某个数据后，还必须通过“刷新”，才能达到修改的目的。

5.2.7　在数据透视图表中创建公式

像其他工作表公式一样，在数据透视表或数据透视图中，可以为计算字段和计算项创建公式，可使用运算符和表达式，还可使用常量并引用报表中的数据，但不能使用单元格

引用和定义的名称，也不能使用以单元格引用或定义的名称为参数的工作表函数，还不能使用数组函数。

在数据透视表或数据透视图中创建计算公式的步骤如下：

①单击数据透视表或数据透视图中数值区中的某一数据。

②选择“数据透视表工具→选项→计算→域、项目和集”，在弹出的菜单中单击“计算字段”（见图 5-25）。

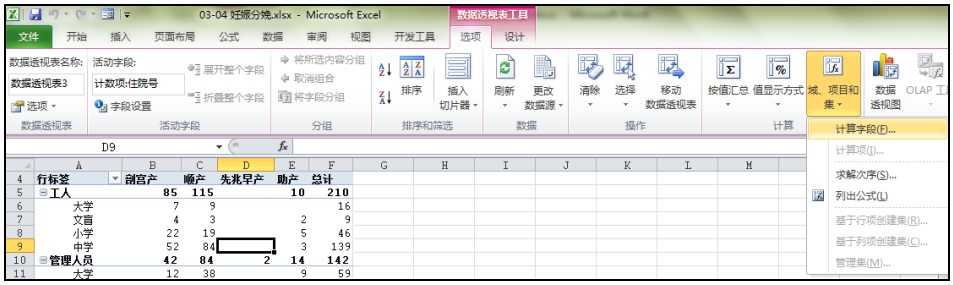


图 5-25　调用“计算字段”

③在“插入计算字段”对话框（见图 5-26）的“名称”框中，键入由公式新创建字段的名称。

④在“公式”框中编辑公式，如“=’体重（kg）’*’身高（cm）’/20”。公式中的体重、身高是通过图 5-26 的字段列表框“插入字段”获得的。

⑤单击“添加”（或修改），再单击“确定”按钮，即可创建新的字段。

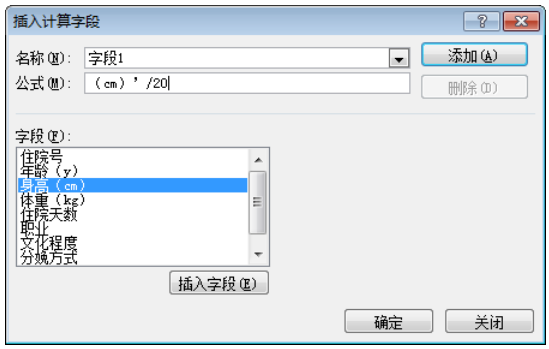


图 5-26　数据透视表中创建新字段

5.3　数据透视图

为了更直观地反映数据透视表的分析结果，可以采用数据透视图绘制统计图形。在绘

制统计图的同时，也会自动产生相应的透视表，并分别存放在不同工作表中。

5.3.1　创建数据透视图

创建数据透视图有两种方式：一是选择“数据透视表工具→选项→工具→数据透视图”；二是在该数据工作表中选取任一单元格，以便 Excel 自动确定用户想用的数据范围，单击“插入→表格→数据透视表→数据透视图”，获得如图 5-27 所示的数据透视图界面。

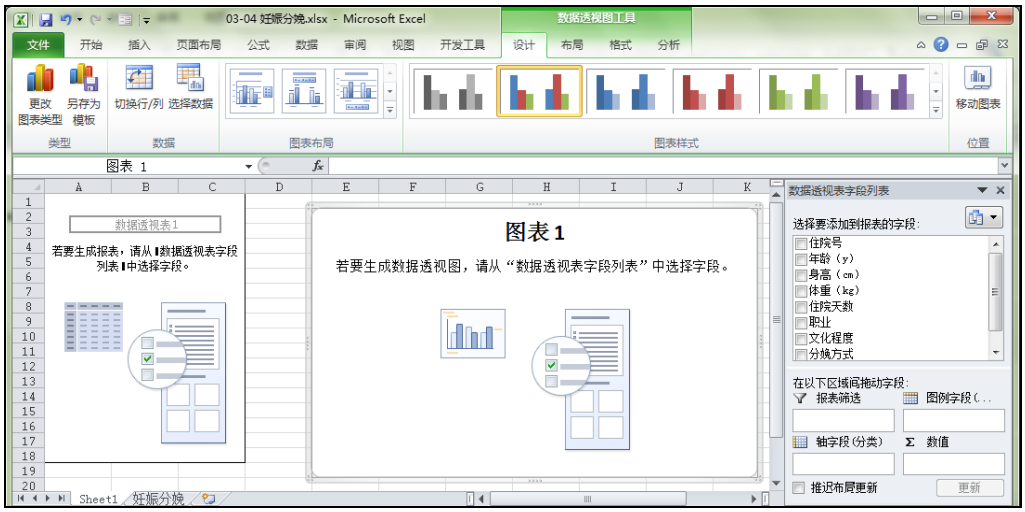


图 5-27　数据透视图雏形初始界面

数据透视图初始界面与数据透视表初始界面（见图 5-8）十分相似，也具有 4 个区域，它们分别是“报表筛选”、“轴字段（分类）”、“图例字段（系列）”和“数值”（见图 5-27）。将数据透视表字段列表中的字段拖放至不同的区域，可获得不同的数据透视图，并自动产生相应的数据透视表。

例 5-8：绘制不同职业中，各种妊娠结局的孕产妇人数构成比饼图。

解：将字段“职业”拖放至“报表筛选”，“妊娠结局”拖放至“轴字段”，“住院号”拖放至“数值”，获得的初始数据透视图如图 5-28 所示。左边是数据透视表，中间是数据透视图，右边上面是字段列表，下方是供拖放、设计透视图或透视表使用的 4 个区域。

5.3.2　调整数据透视图

1. 选择性显示分类变量

初始的数据透视图创建后，可以像数据透视表一样选择分类变量的不同类别。从图

5-28 可见，图左上方的“职业”、右下方的“妊娠结局”之右侧有相应的下拉箭头。为了分类显示不同职业某妊娠结局的数据，可单击相应字段的下拉箭头，去除不需要显示的选项。注意，当只显示了某些类别（不是全部显示所有类别），则会在下拉箭头之后显示出漏斗图案。单击“全部”或选取“全选”，则漏斗图案会消失。

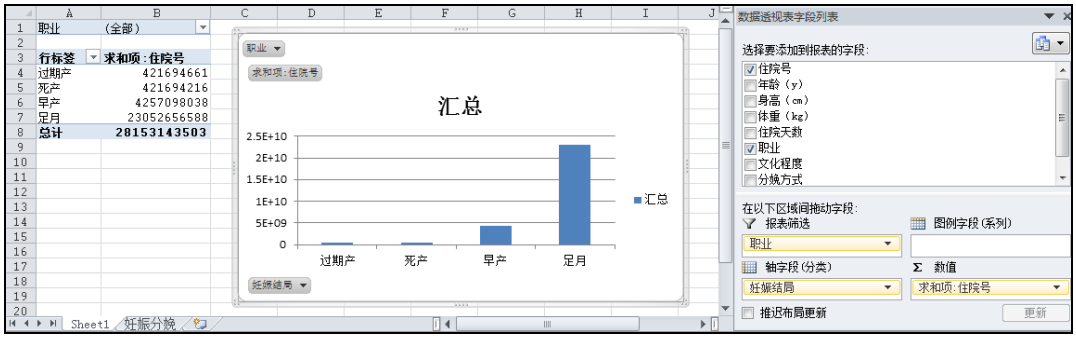


图 5-28　初始数据透视图

2. 设置字段的汇总方式

为了调整字段的汇总方式，单击“数值”区中的“求和项:住院号→值字段设置”，在弹出的“值字段设置”对话框的“值汇总方式”下拉列表中，选择“计数”。这时同一界面的数据透视图（中间）和数据透视表（左侧）都会自动发生变化。

3. 图表类型的选取

在数据透视图中，可利用除 *XY* 散点图、气泡图、股价图外的其他 Excel 图表类型（参见有关统计图制作的章节）。为了显示构成比的情况，可绘制饼图或百分条图。

在绘图区单击鼠标右键，在弹出的快捷菜单中单击“更改图表类型→饼图→分离型三维饼图”，单击“确定”按钮。此时获得的结果如图 5-29 所示。

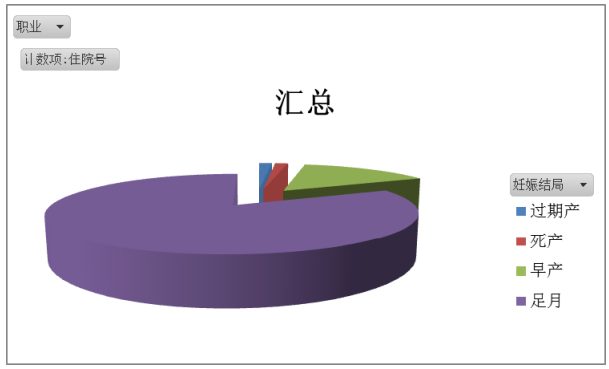


图 5-29　妊娠结局饼图

4. 分页显示图形

如图 5-29 所示是所有职业的妊娠结局构成比，为了分别显示各个职业的妊娠结局构成比情况，只需单击左上角“职业”的右侧下拉箭头，根据需要选择工人、农民、知识分子、商业服务、管理人员等。如图 5-30 是在字段“职业”中选择“知识分子”后获得的图形。

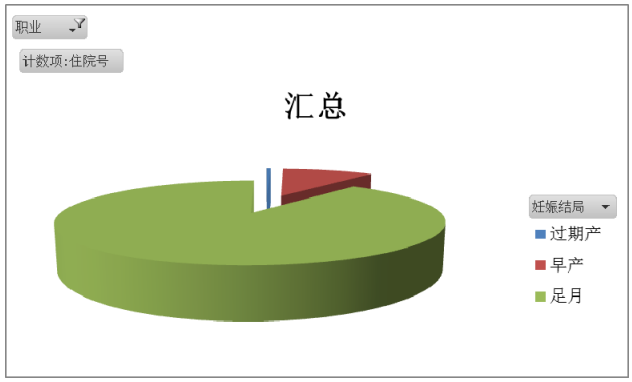


图 5-30　知识分子的妊娠结局饼图

5.3.3　将数据透视图更改为静态图

1. 创建静态图

如图 5-29 和图 5-30 所示的数据透视图中有一般情况下不希望出现的“职业”、“计数项:住院号”、“妊娠结局”等，不符合常用统计图的要求。为了将这种动态图形修改为常用统计图形式，只需在复制粘贴时不选择“图片”格式即可。此时，不希望出现的“职业”、“计数项:住院号”、“妊娠结局”等内容已被隐藏。由图 5-29 改变后的图形如图 5-31 所示。

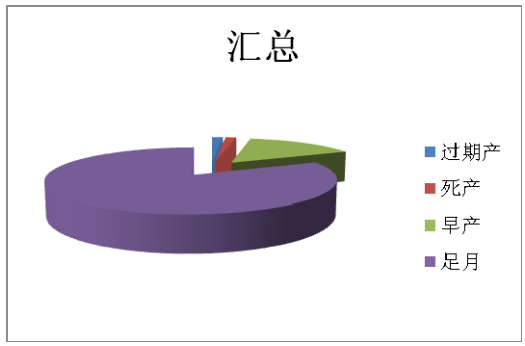


图 5-31　与图 5-29 对应的图形

2. 修饰静态图

为了在图 5-31 中显示每个扇形相应的类别名称与百分比，在绘图区任一扇形上单击鼠标右键，选择“添加数据标签”，然后再次单击鼠标右键选择“设置数据标签格式”。在弹出的“设置数据标签格式”对话框的“标签包括”下去掉“值”并选择“类别名称”、“百分比”，单击“关闭”。

为了增大各扇形间的区分度，可一一修改各扇形的“数据点格式”。比如修改“早产”的“数据点格式”，可按如下操作步骤进行：

两次慢速单击（选取）“早产”扇形（或在其上单击鼠标右键，选择“设置数据点格式”），在“填充”中选取“图案填充”，图案选择第 2 行第 7 个图形（虚线网格），单击“关闭”按钮。其他扇形可仿此修改。删除图例、删除标题，适当调整图形大小等，修饰编辑后的所有职业妊娠结局构成比饼图如图 5-32（也见光盘文件“05-00 妊娠分娩.xlsx”）、图 5-33 所示。

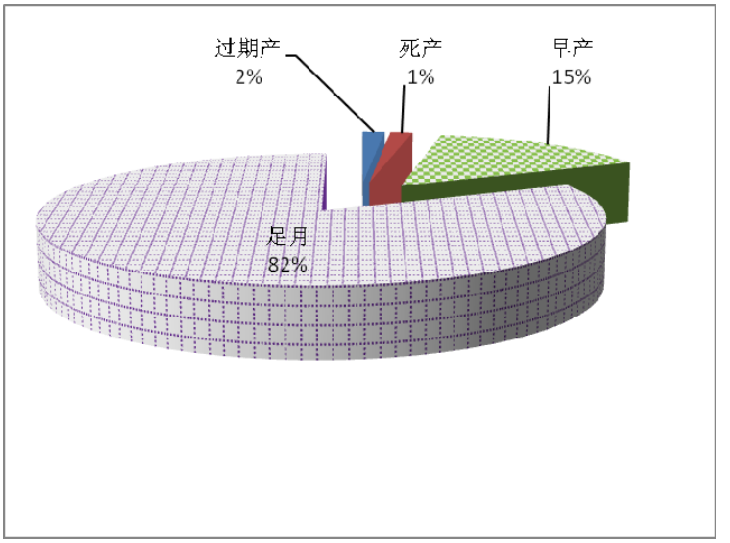


图 5-32　复制粘贴时采用非图片形式后的图形

在数据透视图中，仿照上述办法，可绘制除 *XY* 散点图、气泡图、股价图外的其他 Excel 图表类型。

数据透视图的最大优点是：从原始数据产生图形，而不是利用人工整理的数据；每建立一个数据透视图都将新建一个相应的数据透视表。数据透视表与数据透视图密切关联，当清除数据透视图关联的数据透视表后，数据透视图变成静态图，此时可以像一般情况下一样来编辑绘制的统计图形。

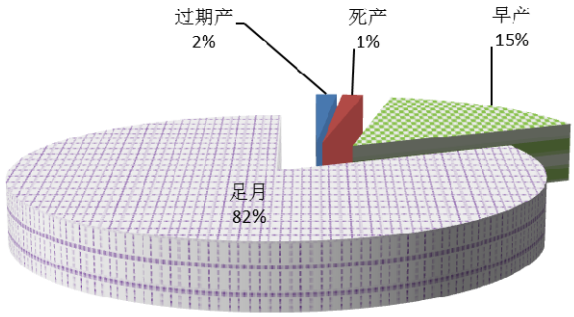


图 5-33　将图 5-32 复制粘贴为图形并去掉图形四周多余部分后的饼图

5.4　数据透视的其他应用

5.4.1　多个工作表的数据透视

在第 4 章 4.8 节中介绍了“多个工作表的合并计算”，对于多个工作表也可采用数据透视分析完成。具体步骤如下：

①打开光盘文件“04-03 多个工作表的合并计算.xlsx”，将鼠标放在工作表“医院 1”的数据区域内任意单元格。在键盘上按顺序按“Alt”、“D”、“P”三个键调出“数据透视表和数据透视图向导”，指定待分析数据源类型为“多重合并计算数据区域”，单击“下一步”按钮。

②指定“自定义页字段”，单击“下一步”按钮。

③单击“选定区域”空白框右侧按钮，将鼠标放在工作表“医院 1”的数据区域内任意单元格，按“Ctrl+A”键选取整个“医院 1”数据区域，单击“选定区域”空白框右侧按钮，单击“添加”按钮，……，直到将工作表“医院 6”添加进入后，单击“下一步”按钮，单击“完成”按钮，即得到与 4.8 节类同的结果（由于年份排序不同，所以显示有所不同。如果将鼠标放在 A5 或其他年份数字单元格，点右键，在弹出的列表中选择“排序”按降序排序，则得到与 4.8 节完全相同的结果）。

5.4.2　数据透视表网页

在 Excel2003 的数据透视表网页内，数据透视表可以像在本机上一样进行字段更改操作与显示。但 Excel2010 已不支持使用 Office Web Components 以交互方式保存 HTML 文件了。这意味着在 Excel2010 中数据透视表网页不能再进行字段更改操作。

例 5-9：光盘文件“05-02 疾病负担.xlsx”是世界卫生组织网上公布的 9 个地区（region）、11 个危险因素（risk factor）、不同性别（sex）、9 组年龄（age）、4 个计算指标（measure）、指标值（value）组成的数据。试以 region、sex 为报表筛选字段，risk factor、measure 为行字段，age 为列字段，对指标值 value 求平均值，并以网页形式显示。

解：数据透视表网页制作步骤如下：

① 在数据区域内选取任一单元格后，以便 Excel 自动确定用户想用的数据范围，单击“插入→表格→数据透视表▼→数据透视表”命令，获得数据透视表初始界面。

② 从数据透视表字段列表中，将“region”、“sex”依次拖放至报表筛选区，“risk factor”、“measure”依次拖放至行标签区，“age”拖放至列标签区，“value”拖放至数值区。

③ 右下角的数值区域单击“求和项:value”，选择“值字段设置”，在弹出的对话框中，“值汇总方式”下拉列表框中选择“平均值”。单击“数字格式”，选择“数值”修改小数位数为“3”，双击“确定”按钮。将“age”“5-14”，通过“移动”方式移至“0-4”之后，将“deaths”移至开头，获得的文件见光盘“05-03 疾病负担透视表.xlsx”。

④ 在功能区中单击“文件→另存为网页”，选择保存类型为“网页（*.htm;*.html）”，修改文件名称为“05-03 疾病负担.htm”，单击“发布”，发布内容选择“数据透视表”，再单击“发布”。这样“疾病负担”数据透视表网页便制作完成（见光盘文件“05-03 疾病负担.htm”，如图 5-34 所示。

region	(全部)									
sex	(全部)									
平均值项:value	列标签									
行标签	0-4	5-14	15-24	25-34	35-44	45-54	55-64	65-74	75+	总计
Alcohol benefit	**0.000**	**0.000**	**-0.150**	**-1.440**	**-4.498**	**-10.884**	**-15.265**	**-29.469**	**-61.730**	**-13.715**
Deaths	0.000	0.000	-0.002	-0.073	-0.267	-0.731	-1.346	-4.035	-18.122	-2.731
DALY	0.000	0.000	-0.299	-2.843	-8.862	-21.403	-29.857	-56.921	-114.399	-26.065
YLD	0.000	0.000	-0.248	-0.960	-2.635	-6.673	-8.127	-11.407	-15.813	-5.096
YLL	0.000	0.000	-0.051	-1.883	-6.228	-14.730	-21.730	-45.514	-98.586	-20.969
Alcohol harm	**1.231**	**2.658**	**58.625**	**35.629**	**46.544**	**36.459**	**29.186**	**26.385**	**20.137**	**28.539**
Deaths	0.045	0.142	1.111	1.369	1.651	2.299	2.434	3.422	5.031	1.945
DALY	2.440	5.245	116.694	70.572	92.263	71.769	57.155	51.059	37.758	56.106
YLD	1.088	1.102	85.674	34.849	53.384	25.035	17.634	12.366	8.184	26.590
YLL	1.353	4.144	31.020	35.723	38.879	46.735	39.521	38.693	29.574	29.516
Cholesterol	**0.000**	**0.000**	**0.019**	**0.706**	**2.940**	**10.441**	**14.678**	**19.104**	**40.536**	**9.825**
Deaths	0.000	0.000	0.001	0.044	0.200	0.715	1.251	2.711	12.208	1.903
DALY	0.000	0.000	0.038	1.390	5.780	20.525	28.731	36.852	74.968	18.698
YLD	0.000	0.000	0.021	0.270	1.146	6.245	8.748	6.617	7.797	3.427
YLL	0.000	0.000	0.016	1.120	4.634	14.280	19.983	30.235	67.171	15.271
Hypertension	**0.000**	**0.000**	**0.051**	**1.096**	**3.442**	**21.155**	**42.681**	**84.742**	**194.611**	**38.642**
Deaths	0.000	0.000	0.001	0.063	0.187	1.312	3.226	10.785	55.199	7.864
DALY	0.000	0.000	0.101	2.160	6.790	41.655	83.749	164.092	361.622	73.352
YLD	0.000	0.000	0.079	0.535	2.378	15.231	31.719	42.921	63.732	17.400
YLL	0.000	0.000	0.022	1.625	4.412	26.423	52.030	121.171	297.890	55.953
Illicit drugs	**0.599**	**0.001**	**35.495**	**28.941**	**19.424**	**2.099**	**1.970**	**1.133**	**1.382**	**10.116**
Deaths	0.026	0.000	0.891	1.561	1.341	0.150	0.164	0.156	0.382	0.519
DALY	1.186	0.002	70.546	57.102	38.177	4.117	3.858	2.188	2.574	19.972
YLD	0.387	0.002	45.790	16.339	6.317	1.032	1.161	0.322	0.002	7.928
YLL	0.798	0.000	24.755	40.763	31.860	3.096	2.697	1.866	2.572	12.045

图 5-34　数据透视网页

统计分析篇

第 6 章　统计图表

第 7 章　统计描述

第 8 章　区间数据的假设检验

第 9 章　名义与有序数据的假设检验

第 10 章　相关与回归

第 6 章

统 计 图 表

统计图表是统计分析的基础，它看似简单，但实际工作中出错率仍然很高。统计分析的内容用表格列出，称为统计表（statistical table），可概括性地展示数量间的关系，便于计算、分析和对比。统计图（statistical graph）是用点的位置、曲线的走势、直条的长短或面积的大小等形式，直观地呈现所研究事物的数量关系。统计表和统计图在揭示资料所蕴涵的信息方面各有千秋，前者数量精准，后者形象直观，有时二者结合使用。不过，为了节省篇幅，期刊杂志中用了统计表，通常不再用统计图重复表达，反之亦然。

本章先简要介绍统计表的制作，然后重点介绍统计图的制作。不同的期刊杂志可能有不同的统计图表制作要求，读者应在本章的基础上参照各期刊杂志要求进行修改。

6.1　统计表的制作

统计表由标题、标目、线条、数字、注释等组成（见表 6-1 及图 6-1）。

表 6-1　食物中毒患者潜伏期的频率分布

潜伏期（h） （1）	频率，f（%） （2）	累计频率，$\sum f$（%）* （3）	
0～	14.91	14.91	
12～	36.73	51.64	
24～	23.27	74.91	
36～	14.18	89.09	
48～	6.91	96.00	
60～	3.27	99.27	
72～	0.73	100.00	



图 6-1　统计表的结构

统计学对统计表的一般要求是：

①**标题**应概括性地阐明表的内容，必要时注明研究开展的时间和地点，标出本表在全文中的序号（见表 6-1），写于表的上方。

②**标目**有横标目和纵标目之分，标目应文字简明，如有量纲则需注明量纲，如表 6-1 中的频率和累计频率下面的数字是百分数，应标注“%”。图 6-1 中的第 C 列为纵标目（通常为分类变量，充当主语），第 3 行是横标目（通常为统计指标，充当谓语）。与“数字”一起从左到右、从上到下可以读成一句话。

③**线条**不宜过多，除顶线、底线以及纵标目下面与合计上面的横线外，其余尽量省去，

不使用竖线和斜线。

④**数字**（见图 6-1 中的 Excel 单元格区域 E4:G10）尽可能采用阿拉伯数字，同类指标的小数位数一致，纵向位次对齐，如小数位对小数位，个位对个位。表内不留空白，有尚未观察的内容可用“…”表示；不存在的内容可用“—”表示；观察结果为“0”，需写出“0”。

⑤表中需要**注释**的地方标出“*”或其他符号，具体解释内容写在表的下方（见图 6-1）。



注意：制作表格时，度量衡单位或列标号（见表 6-1）通常需要在一个单元格内转行显示，这种情况下可以在 Excel 活动单元格或编辑栏采用 Alt+Enter 组合键实现。

6.2 统计图的结构及特点

由图 6-2（见光盘文件“06-01 统计图结构.xlsx”）可见，一张统计图的整个区域为图表区；在图表区中由坐标轴围成的区域为绘图区；直条、线条、扇面、圆点等就是数据（系列）标志；若干数据坐标点组成一个数据系列，一般情况下 Excel 图形只有一个数据系列，实际上在一个绘图区可以采用多个数据系列绘制图形。坐标轴一般有纵、横两个轴。横轴也叫 *X* 轴，一般为分类轴，但 *XY* 散点图等图形的 *X* 轴却是数值轴；纵轴也叫 *Y* 轴，或数值（*Y*）轴。三维图表有第 3 个轴（*Z* 轴）。饼图、圆环图没有坐标轴。坐标轴上有刻度线，刻度线对应的数字叫刻度线标签。坐标轴旁有坐标轴标题。统计图标题一般应位于绘图区的下方，用来说明图表的内容。如果用不同的颜色或图案反映或说明不同的事物，则需要采用图例。图例可根据图表的情况放在方便、恰当的位置。

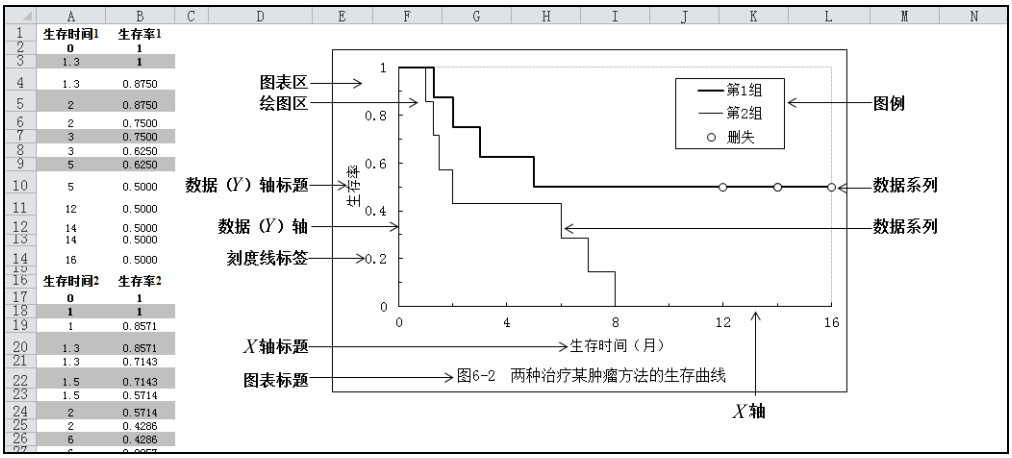


图 6-2 统计图的一般结构

Excel 绘制图形有如下重要特点。

①功能强大。Excel 提供了 11 种标准图表类型（参见图 6-3），每种图表类型中含有 2～19 种子图表。

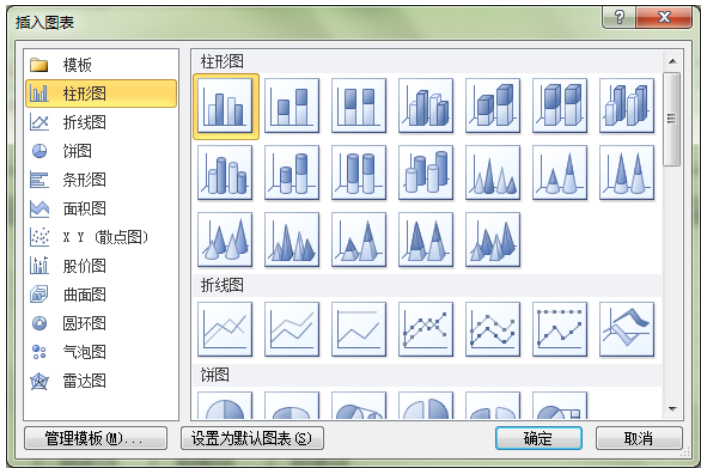


图 6-3 “插入图表”对话框

②修饰后的图形美观好看。用户可以对图表区、绘图区、数据系列、坐标轴（标题）、图例、图表标题的格式，如文字的颜色、字体、大小，以及图背景图案、颜色等进行修改、调整，使图表赏心悦目。

③图与数据关系密切，具有“即改即可见”的动态效果。当图和数据放在一张工作表上时，改变产生图的数据，其图形将发生相应变化；将鼠标放在图中某数据点上，在鼠标下方将弹出一个文本框给出数据点的系列名称及具体数据坐标值；用鼠标单击绘图区中的“数据系列”标志，其图所属数据单元格（见图 6-2 的左侧第 A、B 列）将被彩色框线围住，便于用户查看图形的数据引用位置。

④ Excel 统计图也可实现统计功能。在“数据系列”上单击鼠标右键，可以向散点图、线图、条图等添加趋势线，并可给出趋势线的回归方程与决定系数。

6.3　Excel 标准图表类型简介

Excel 标准图形共有 11 个，简介如下。

（1）柱形图

柱形（column）图是用户经常使用的一种图表类型。通常采用柱的长度来描述不同类别之间数据的差异。柱形图共有 19 种子图表类型：簇状柱形（圆柱、圆锥、棱锥）图、堆积柱形（圆柱、圆锥、棱锥）图、百分比堆积柱形（圆柱、圆锥、棱锥）图、三维簇状柱

形（圆柱、圆锥、棱锥）图、三维堆积柱形图、三维百分比堆积柱形图和三维柱形图等。其中堆积柱形图和百分比堆积柱形图通过将不同类别数据堆积起来反映相应的数据占总数的大小。除簇状柱形图、堆积柱形图、百分比堆积柱形图外，其他均具有立体感，进一步加强了修饰效果。这些图形可用来绘制单式条图、复式条图、分段条图、直方图、百分条图等。注意：一般情况下，数值轴的起点应该为“0”。

（2）折线图

折线（line）图是用直线段将各数据点连接起来而组成的图形，以折线方式显示数据的变化趋势。折线图常用来分析数据随时间的变化趋势，也可用来分析比较多组数据随时间变化的趋势。在折线图中，一般情况下水平轴（*X* 轴）用来表示时间的推移，并且时间间隔相同；而垂直轴（*Y* 轴）代表不同时刻的数值的大小。折线图共有 7 个子图表类型：折线图、堆积折线图、百分比堆积折线图、数据点折线图、堆积数据点折线图、百分比堆积数据点折线图和三维折线图。

（3）饼图

饼（pie）图是用圆的总面积（100%）表示事物的全部，用各扇形的面积表示各个组成部分，各组成部分的面积之和为 1，即 100%。通常用饼图描述百分比构成情况。饼图共有 6 个子图表类型：饼图、三维饼图、复合饼图、分离型饼图、分离型三维饼图和复合条饼图。其中复合饼图和复合条饼图是在主饼图的一侧生成一个较小的饼图或堆积条形图，用来进一步反映某一扇形内部各组成部分的构成比。

（4）条形图

条形图（bar）与柱形图十分相似。用等宽直条的长短表示各指标数量的大小。所比较的数值可以是绝对数，也可以是相对数。条形图共有 15 个子图表类型：簇状条形图、堆积条形图、百分比堆积条形图、三维簇状条形图、三维堆积条形图、三维百分比堆积条形图、簇状水平圆柱（圆锥、棱锥）图、堆积水平圆柱（圆锥、棱锥）图、百分比堆积水平圆柱（圆锥、棱锥）图。

（5）面积图

面积（area）图实际上是折线图的另一种表现形式，它使用折线和分类轴（*X* 轴）组成的面积以及两条折线之间的面积来显示数据系列的值。面积图除了具备折线图的特点，强调数据随时间的变化以外，还可通过显示数据的面积来分析部分与整体的关系。面积图共有 6 个子图表类型：面积图、堆积面积图、百分比堆积面积图、三维面积图、三维堆积面积图和三维百分比堆积面积图。

（6）*XY* 散点图

XY 散点（scatter）图与折线图看似类似，但它们之间有较大差别。其不同之处在于：折线图要求横轴等间距，且数据为分类而不是数值，即使为数值，它也当作文本看待；而

XY 散点图的横轴为数值轴，可以不等间距。散点图不仅可以用直线段反映时间的变化趋势，而且可以用光滑曲线或一系列散点来描述数据。*XY* 散点图除了可以显示数据的变化趋势以外，更多地用来描述数据之间的关系。例如，两组数据之间是否相关，是正相关还是负相关，以及数据之间的集中趋势或离散趋势情况等。*XY* 散点图共有 5 个子图表类型：散点图、平滑线散点图、无数据点平滑线散点图、折线散点图和无数据点折线图。

（7）股价图

股价（stock）图是一类比较复杂的专用图形，通常需要特定的几组数据。主要用来判断股票或期货市场的行情，描述一段时间内股票或期货的价格变化情况。股价图共有 4 种子图表类型：盘高—盘低—收盘图、开盘—盘高—盘低—收盘图、成交量—盘高—盘低—收盘图和成交量—开盘—盘高—盘低—收盘图。其中的开盘—盘高—盘低—收盘图也称 K 线图，是股市上判断股票行情最常用的技术分析工具之一。

（8）曲面图

曲面图（surface）是折线图和面积图的另一种形式，它在原始数据的基础上，通过跨两维的趋势线描述数据的变化趋势。此外，可通过拖放曲面图的坐标轴，方便地从不同角度变换、观察数据。曲面图的子图表类型有：三维曲面图、三维曲面框架图、曲面图和曲面俯视框架图。

（9）圆环图

圆环（doughnut）图的用途与饼图相似。

（10）气泡图

气泡图（bubble）是 *XY* 散点图的扩展，它相当于在 *XY* 散点图的基础上增加了第 3 个变量，即气泡的大小尺寸。气泡图含有两个子图表类型：气泡图和三维气泡图。

当有两列数据时，第 1 列数据将反映 *Y* 轴的值，第 2 列数据将反映气泡的大小；当有 3 列数据时，第 1 列数据将反映 *X* 轴的值，第 2 列数据将反映 *Y* 轴的值，第 3 列数据将反映气泡的大小。

（11）雷达图

雷达（radar）图由一个中心向四周射出多条数值坐标轴。如果比较多个数据系列指标，则每个指标都拥有自己的数值坐标轴，把每个数值坐标轴上同一数据序列的值用折线连接起来便形成雷达状图形。雷达图常用来比较拥有若干个数据序列的指标。雷达图的 3 个子图表类型分别为雷达图、数据点雷达图和填充雷达图。

6.4　创建统计图的一般步骤

下面采用复式条图的例子，详细说明 Excel 创建统计图的方法。

例 6-1：2000 年甲、乙、丙三地年报表资料显示：卡介苗、脊髓灰质炎疫苗、百白破疫苗、麻疹疫苗（简称四苗）的接种率（%），如图 6-4（见光盘文件“06-02 复式条图.xlsx”）所示，请根据左侧表中数据创建统计图，以此比较三地的四苗接种率情况。

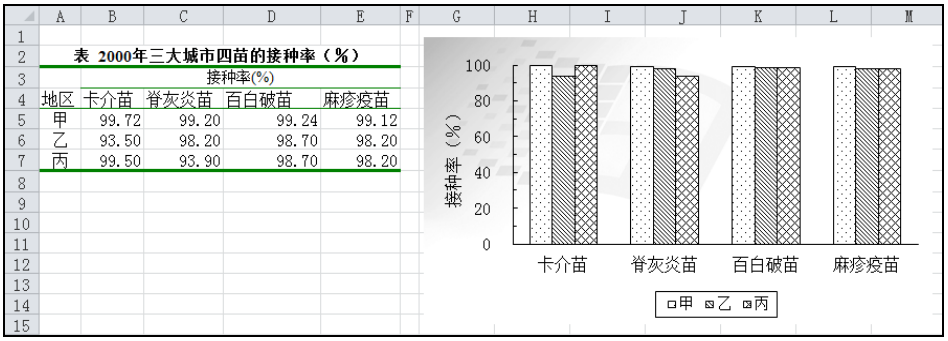


图 6-4　2000 年甲、乙、丙三大城市四苗接种率（%）

解：Excel创建统计图按下列步骤进行。

①在表格区外任意位置放置光标，单击“插入→图表→柱形图▼→簇状柱形图”，见图 6-5 所示（将鼠标指针停留在任何图表类型上时，屏幕提示将会显示其名称）。若要查看所有可用的图表类型，请单击“图表”命令组的右下角按钮，以启动“插入图表”对话框，这里所有图表类型都在其中，如图 6-3 所示。

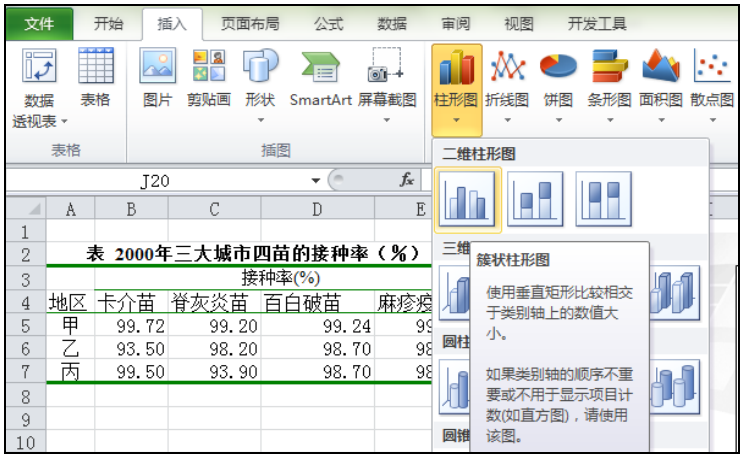


图 6-5　插入簇状柱形图

②单击“图表工具→设计→选择数据”，见图 6-6（**提示：**若“图表工具”选项卡不可见，请单击所生成图形的任何位置将其激活），将弹出“选择数据源”对话框。单击“图表数据区域”右侧的参数框，选取单元格区域 A4:E7，即出现图 6-7 的界面。

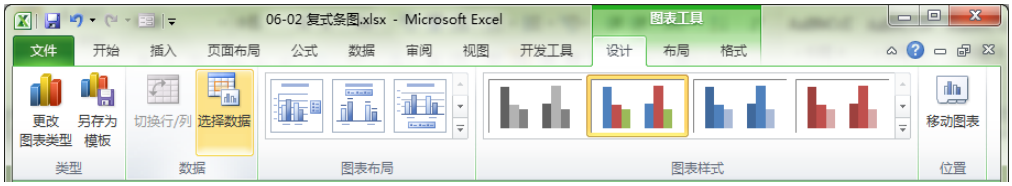


图 6-6　“选择数据”按钮的位置

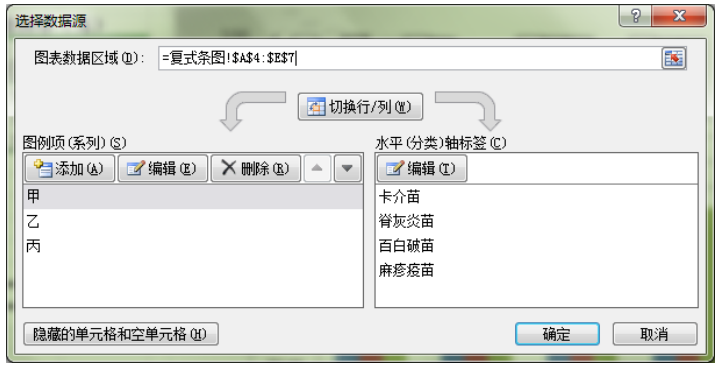


图 6-7　“选择数据源”对话框

③ 单击“确定”按钮，即获得初始的统计图，如图 6-8 所示。

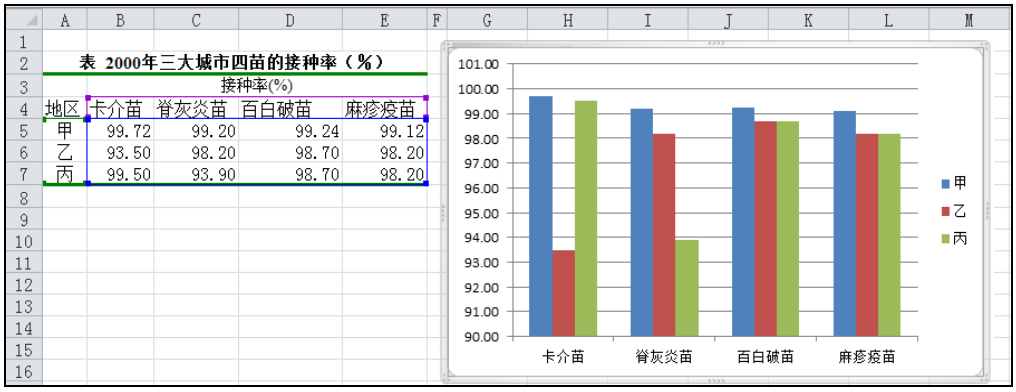


图 6-8　统计图初始界面

图 6-8 符合比较甲、乙、丙三地的四苗接种率的含义（如果需要比较四苗接种率分别在甲、乙、丙三地的情况，则应在“选择数据源”对话框中单击“切换行/列”按钮，或者单击“图表工具→设计→数据→切换行/列”按钮）。从图 6-8 可见，新创建的统计图与图 6-4 有所不同，数值轴起点不为 0（设置不正确）。那么，如何使图 6-8 变成图 6-4 的形式呢？需要采用下一节“统计图的编辑与修饰”的方法对图形进行美化。

6.5 统计图的编辑与修饰

统计图的编辑与修饰需要使用“图表工具”选项卡。当选取刚创建的统计图时，会在 Excel 功能区出现“图表工具”选项卡，“图表工具”下又分为设计、布局和格式 3 个选项。

“设计”选项卡：应用于整体统计图的修改。如更改统计图的图表类型（“类型”），更改统计图的数据来源（“数据”），改变统计图整体布局（“图表布局”），改变统计图整体颜色（“图表样式”），移动整个统计图所在工作表到不同工作表中（“位置”）等。

“布局”选项卡：应用于调整图表的布局，显示不显示图表中某些元素。这是编辑与修饰统计图形最重要的选项之一。可以进行标题、标签修改（“标签”组），坐标轴、网格修改（“坐标轴”组），背景色彩、三维旋转修改（“背景”），趋势线、误差线的计算与分析（“分析”组），图中插入图片、图形、文本（“插入”）等。在“布局”选项卡的最左侧，还有“当前所选内容”组，通过单击该组中的“设置所选内容格式”按钮，还可对所在当前位置（图表区、绘图区、数据系列、坐标轴、图例）内容的格式进行修改，如图 6-9 所示。



图 6-9 “图表工具”下“布局”选项卡的界面

“格式”选项卡：修改图表中的其他格式。

下面主要针对如图 6-8 所示的统计图进行编辑与修饰。

（1）删除网格线与背景色

一般的科研统计图不需要网格线，可以将其去除。去除网格线有两种方法：一是在任意一条网格线上单击鼠标右键，在快捷菜单中选择“删除”命令；二是在单击“图表工具→布局→坐标轴→网格线▼→主要横网格线▼/主要纵网格线▼→无”。如果需要显示主要（次要）网格线，则可单击“图表工具→布局→坐标轴→网格线▼→主要横网格线▼/主要纵网格线▼”，在其下拉列表中设置“主要网格线”与“次要网格线”。

（2）修改横坐标轴格式

从图 6-8 中可见，横坐标上具有不希望出现的刻度线。为了去除刻度线，将光标放在分类（X）轴上，单击鼠标右键，在快捷菜单中选择“设置坐标轴格式→坐标轴选项”，

将主要（次要）刻度线选择为“无”即可去除刻度线（见图 6-10）。

（3）修改纵坐标轴格式

图 6-8 中纵坐标的最小值不为 0，这违反了条图的数值轴起始点应为 0 的原则。为了修改，将鼠标移至数值（*Y*）轴，单击鼠标右键，选择“设置坐标轴格式→坐标轴选项”，将“最小值”、“最大值”、“主要刻度单位”、“次要刻度单位”、“坐标轴值”等文本框内的数值分别修改为 0、100、20、10、0（见图 6-11），并在该对话框中将主要及次要刻度线类型设置为内部，以便在轴内侧显示刻度线。

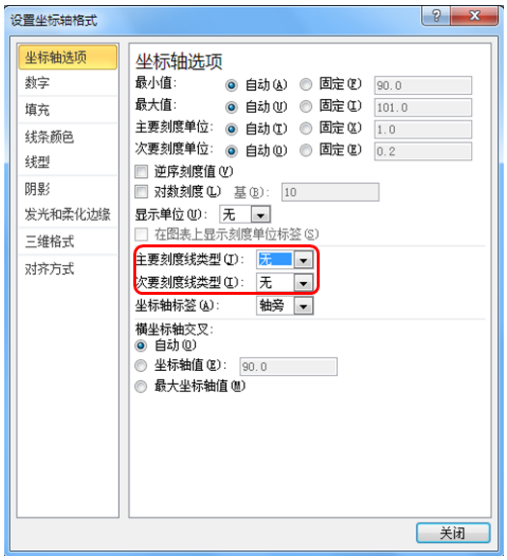


图 6-10　去除刻度线

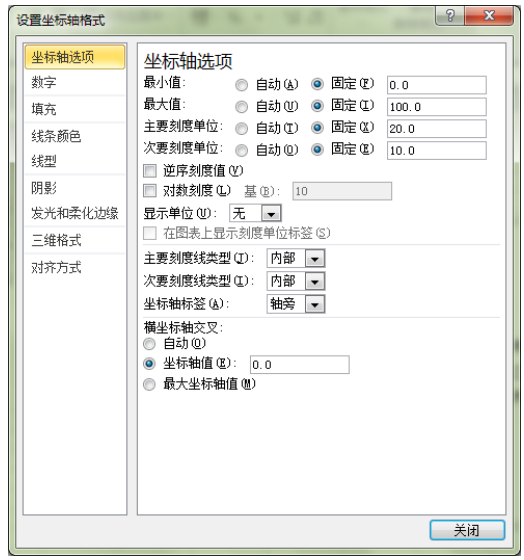


图 6-11　“设置坐标轴格式”对话框

要增加或减少坐标轴刻度线标志数据的小数位数，可在“设置坐标轴格式”对话框中选择“数字→类别→数字”选项卡，将小数位数设为 0。

（4）修改数据系列格式

很多情况下，科研杂志要求采用黑白图片，这就需要对彩色图片采用图案的填充效果进行修改。具体操作如下所示。

①将鼠标移至其中一个数据系列的标志图（如数据系列“甲”）上，单击鼠标右键，此时该数据系列被全部选中，并弹出一个快捷菜单，选择“设置数据系列格式”命令，弹出“设置数据系列格式”对话框（见图 6-12）。可设置系列选项、填充、边框颜色、边框样式、阴影、发光和柔化边缘及三维格式等。

②单击右方的“填充”选项卡，将弹出“填充”界面（注意：该对话框也可在图表区格式、绘图区格式中出现，其设置方法与数据系列格式的设置完全相同）。对话框中有“无

填充”“纯色填充”“渐变填充”“图片或纹理填充”“图案填充”“自动”和“以互补色代表负值”等选项。“图案填充”中，Excel 2010 有 48 种基本图案可供选择，通过改变前景色与背景色，又可组合出很多图案。本例先选择前景色为黑色、背景色为白色，并选中第一个图案，双击“确定”按钮获得甲系列的图案。再单击“边框颜色”选项卡，选择实线、黑色来设置边框。“乙”、“丙”系列以同样方式进行类似设置。

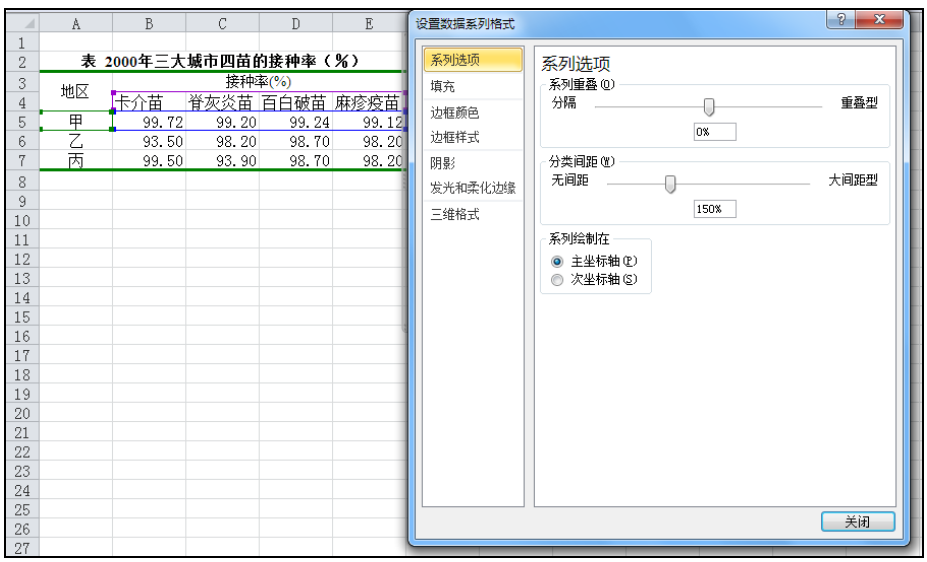


图 6-12　设置数据系列格式

（5）图例与坐标轴标题的编辑

在图例上右击，选择“设置图例格式→图例选项→图例位置→底部”。图例与图重叠的话，将图上移。单击图例，图例周围会出现 8 个点，鼠标放在点上会显示出双箭头，向各个方向拖动可改变图例框的形状与排列形式。此外，还可对图例的其他格式（如图案、字体）做出修改。

单击“图表工具→布局→坐标轴标题→主要纵坐标轴标题→旋转过的标题”，在出现的文本框中输入标题“接种率（%）”，去掉加粗格式，完成坐标轴标题设置。

（6）图表区格式的编辑

单击图表区，图表区周围呈现边框线，将鼠标放在 4 个角或每个边框线的中间，当十字箭头✣ 改变为双向箭头时，向各个方向拖动可改变整个图形的形状与大小；在此情况下，也可对整个表格的文字大小、字体等格式进行设定。

在图表区单击鼠标右键，在其快捷菜单中选择“设置图表区域格式”命令，将弹出“设置图表区格式”对话框，该对话框格式与“设置绘图区格式”对话框、“设置数据系列格

式”对话框相似，所以该对话框的设置与它们也相似。图表区、绘图区、数据系列均可设置其填充效果，“填充”下方均有“无填充”、“纯色填充”、“渐变填充”、“图片或纹理填充”、“图案填充”和“自动”选项，其中图片填充可以是各种电子图片。图 6-4（也见光盘文件“06-02 复式条图.xlsx”）是将微软 Windows 标识图片放在图表区的效果，也可以将图片放在绘图区或数据系列上，读者不妨自己试一试。

在设置图表区域格式时，可设置图表区的边框颜色，为了去除边框线，本例设置为无线条。

（7）向其他图表类型的转化

图形绘制出来后，仍可改变图表类型、数据来源、图表选项、图表位置等。在图表区、绘图区、数据系列标志等位置单击鼠标右键，在不同的快捷菜单中均可出现图表类型。

对于图 6-4，在图表区或绘图区单击鼠标右键，选择“更改图表类型”命令，在对话框中选择“柱形图→簇状圆柱图”选项，单击“确定”按钮。再单击“图表工具→布局→三维旋转”按钮，进行适当修改，可获得如图 6-13（也见光盘文件“06-03 复式条图（三维图）.xlsx”）所示的效果。

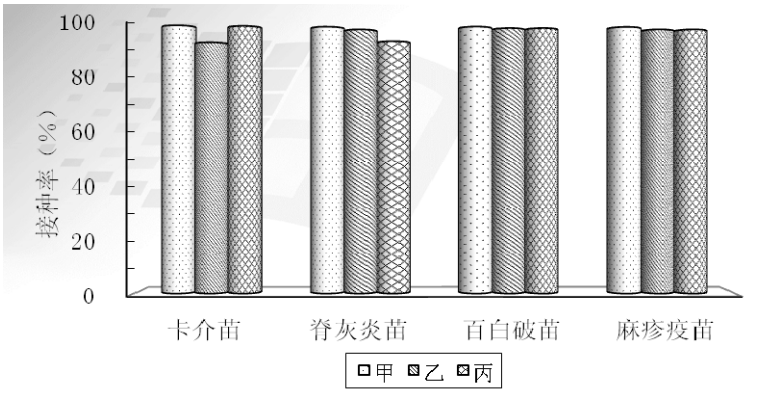


图 6-13　簇状圆柱图

6.6　绘制统计图举例

6.6.1　分段条图

条图（bar chart）是用等宽直条的长短表示统计指标数量的大小，适用于彼此独立的统计量指标间的比较。统计指标可以是例数、率、平均数等。一般分为单式条图和复式条图。

绘制条图应注意：

①数值（*Y*）轴的刻度必须从0开始，否则会改变各个对比组之间的比例关系。

②为了便于统计图的查看与比较，分类（*X*）轴各直条一般按统计指标的大小顺序或其自然顺序排列。

③各直条的宽度相等，间隔的宽度一般与直条的宽度相等，或为直条宽度的一半。

前面的例 6-1 绘制的就是一种条图，这种条图也称为复式条图，其中有两个分类指标（甲、乙、丙 3 个地区、4 种疫苗）、一个统计指标（接种率）。当分类指标只有一个时称单式条图。当有一个分类指标、两个或两个以上统计指标时，宜绘制分段条图。

例 6-2：某地选取老、中、青三代人各若干，获得其结核菌素阳性率分别为 50.28%、41.23%和 30.10%，其中强阳性率分别为 12.65%、5.87%、2.01%。请绘制一张统计图比较老、中、青三代人的结核菌素阳性率。

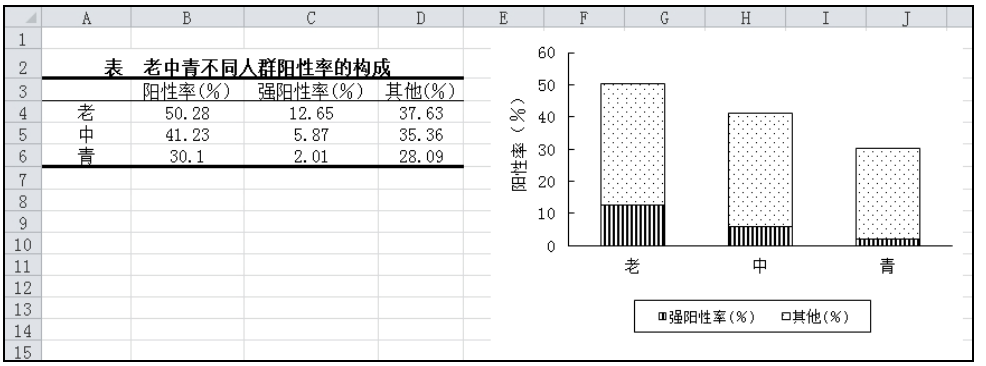


图 6-14　某年某地三代人的结核菌素阳性率

解：本例强阳性率为阳性率的一部分，因此可以将阳性率分解为强阳性率与其他两个组成部分，整个直条代表阳性率。该资料宜绘制分段条图（component barchart），为了绘制出图 6-14（也见光盘文件“06-04 分段条图.xlsx”）右侧的图形，可按如下步骤进行。

① 按图 6-14 输入第 A、B、C 3 列文字或已知数据，在单元格 D3 输入“其他（%）”，第 D 列数据由第 B 列减第 C 列获得，即在单元格 D4 输入公式“=B4–C4”，并将此公式向下拖放至单元格 D6。

② 选取单元格区域 C3:D6，单击“插入→柱形图→堆积柱形图（stacked column，柱形图中第 2 个子类型）”（见图 6-5）。

③ 在网格线上单击鼠标右键，选择清除网格线；在绘图区或图表区单击鼠标右键，选择快捷菜单中的“选择数据”命令，在“选择数据源”对话框中，单击“水平（分类）轴标签”下面的“编辑”按钮，用鼠标选取单元格区域 A4:A6，在弹出的“轴标签”对话框

中选择“确定”，再单击“确定”选项，修改横坐标标题；用鼠标右键单击分类轴，在“设置坐标轴格式”对话框中的“坐标轴选项”中设置主要和次要刻度线类型均为无。

④在某一系列数据（直条）标志上单击鼠标右键，选择“设置数据系列格式→填充→图案或纹理填充”单选项，适当选择前景色、背景色与图案；并在“设置数据系列格式→边框颜色”中设置边框为实线，颜色为黑色，单击“关闭”；另一系列数据（直条）标志仿此设置；在“设置数据系列格式→系列选项→分类间距”中调整间距为 100%。

⑤在纵轴上单击鼠标右键，选择“设置坐标轴格式→数字→类别→数字”选项卡，小数位数调整为 0，单击“关闭”。

⑥在功能区“图表工具→布局→坐标轴标题→主要纵坐标轴标题→旋转过的标题”选项，设置纵坐标轴标题为阳性率%。适当调整图例的位置、形状，调整绘图区的大小，调整整个图表区域的字体大小等，可获得图 6-14 右侧的分段条图。

6.6.2　误差条图

误差条图（error-bar）是用条图或线图表示均数的基础上，在图中附上标准差的范围，以此反映数据个体值离散情况的一种统计图。

例 6-3：4 种营养素喂养小白鼠 3 周后所增体重（克）的均数标准差数据见图 6-15（也见光盘文件“06-05 误差条图.xlsx”）左侧，试制作误差条图。

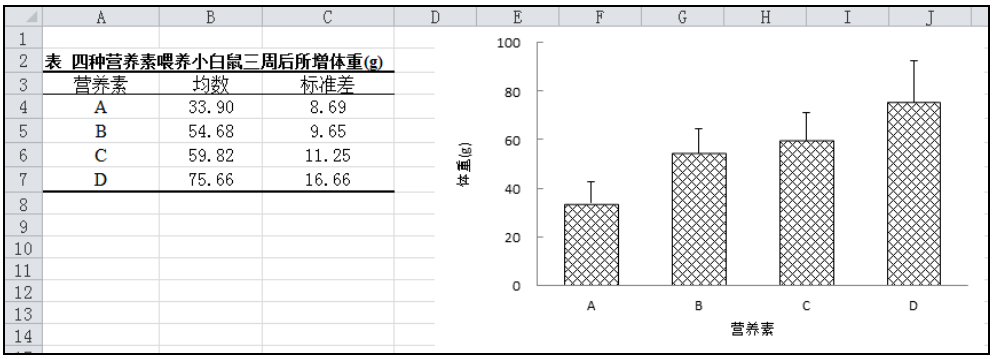


表 四种营养素喂养小白鼠三周后所增体重(g)		
营养素	均数	标准差
A	33.90	8.69
B	54.68	9.65
C	59.82	11.25
D	75.66	16.66

图 6-15　4 种营养素喂养小白鼠 3 周后所增体重

解：制作步骤如下。

①按图 6-15 输入第 A、B、C 三列文字或已知数据。

②选取单元格区域 A3:B7，单击“插入→柱形图→簇状柱形图（柱形图中第 1 个子类型）”，见图 6-5，单击“图表工具→布局→坐标轴标题→主要横坐标标题→坐标轴下方标题”，在标题框中填入“营养素”，单击“图表工具→布局→坐标轴标题→主要纵坐标标

题→旋转过的标题”，在标题框中填入“体重（g）”。

③去除“均数”二字；在网格线上单击鼠标右键，选择“删除”。

④在数据系列（直条）上单击鼠标右键，在“设置数据系列格式→填充→图案填充”，适当选择前景色、背景色与图案（倒数第 2 行最后一个）；并在“设置数据系列格式→边框颜色”中设置边框为实线，颜色为黑色，单击“关闭”按钮；在“设置数据系列格式→系列选项→分类间距”中调整间距为 100%；选择“图表工具→布局→分析→误差线→其他误差线选项”命令，弹出“设置误差线格式”对话框（如图 6-16 所示），单击对话框左侧“垂直误差线”选项卡后，在“显示”栏中选择“正偏差”和“线端”单选项，在“误差量”栏中选择“自定义”单选项。在弹出的“自定义错误栏”对话框中选取“正错误值”单元格区域 C4:C7（即标准差），依次单击“确定”、“关闭”按钮，此时标准差即在图形中显示出来。

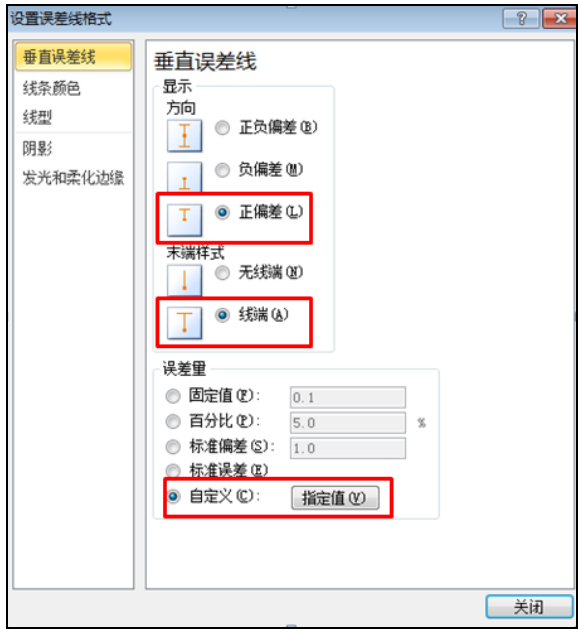


图 6-16 “设置误差线格式”对话框

注意：有时误差线并非理想情况，此时可在其中任意一条误差线上单击鼠标右键，选择“设置错误栏格式”命令，打开“设置误差线格式”对话框（见图 6-16），可根据提示进行显示方式和误差量调整。值得注意的是，“垂直误差线”下方还有“线条颜色”、“线型”、“阴影”、“发光和柔化边缘”选项卡，可根据需要进行选用。

⑤在纵轴及其标题区内单击鼠标右键，选择“设置坐标轴格式→数字→类别→数字”选项，小数位数调整为 0，单击“关闭”按钮。

⑥适当调整绘图区的大小，调整整个图表区域的字体大小等，并进行其他必要的修改，可获得图 6-15 右侧的误差条图。

⑦复制图 6-15，在新图的图表区单击鼠标右键，在其快捷菜单中选择“选择数据”命令，在弹出的对话框中单击“图例项（系列）”下的“添加”按钮，在弹出的编辑数“系列值”下设置单元格区域 B4:B7，单击“确定”按钮；分类轴标签选取 A4:A7（默认），单击“确定”按钮。在新产生的数据系列（直条）上单击鼠标右键，在其快捷菜单中选择“更改系列图表类型→折线图（折线图第 1 个子类型）”，单击“确定”按钮，此时各直条顶部用直线已连接起来。在该直线上单击鼠标右键，在弹出的快捷菜单中选择“设置数据系列格式”命令，将线型样式改为虚线，颜色改为黑色，单击“确定”按钮即可获得如图 6-17 所示的效果。

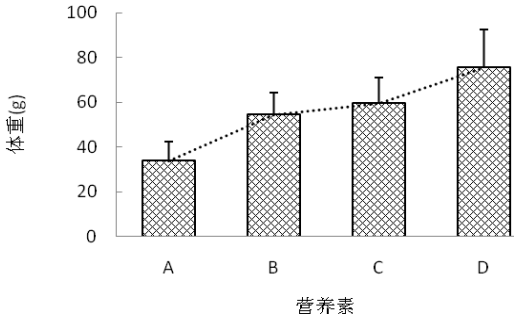


图 6-17　带有线条连线的误差条图

6.6.3　百分条图

百分条图（percentage chart）是用一个矩形条的总面积（100%）表示事物的全部，将此矩形划分成若干个小的矩形，每个小矩形的面积表示各个组成部分。各组成部分的面积之和为 1，即 100%。在一个百分条图中可以有多个矩形条表示多组构成比数据。

例 6-4：胃癌按病理组织学可划分为分化癌、低分化腺癌、未分化癌。某地这几种类型男性的构成比分别为 52.3%、27.8%、19.9%，女性的构成比分别为 30.2%、18.1%、51.7%。试比较该地区男、女胃癌的类型有何不同。

解：此资料需要绘制成如图 6-18（也见光盘文件“06-06 百分条图.xlsx”）右侧所示的图形。绘制该图的步骤如下。

①按图 6-18 输入第 A、B、C、D 4 列文字或已知数据，在单元格 E3 中输入“合计”。

在单元格 E4 中输入公式“=SUM(B4:D4)”（也可在“公式”选项卡中单击自动求和图标 Σ），并将此公式向下拖放至单元格 E5。

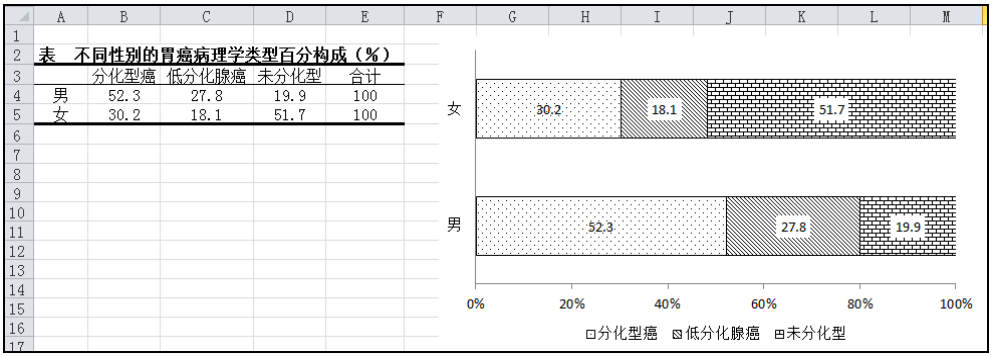


图 6-18　某地区男、女胃癌患者病理组织学类型

② 选取单元格区域 A3:D5，单击“插入→图表→条形图→百分比堆积条形图（二维条形图中的第 3 个子类型）”，在“图表工具→设计→切换行/列”。

③ 在网格线上单击鼠标右键，选择“删除”；用鼠标右键单击分类轴（实际是垂直的轴），在“设置坐标轴格式→坐标轴选项”，选择主要和次要刻度线类型均为无。

④ 在某一系列数据（矩形条）上单击鼠标右键，选择“设置数据系列格式→填充→图案填充”，适当选择前景色、背景色与图案，单击“确定”；另两个系列数据（矩形条）仿此设置；在“设置数据系列格式→系列选项”中调整间距宽度与直条宽度相等。

⑤ 单击“图表工具→布局→标签→数据标签→居中”，可在每个数据系列的中间显示该部分所占百分数。由于“低分化腺癌”、“未分化型”的图案太深，需要右击这些图案中的数据标签，来“设置数据标签格式”，单击“填充→纯色填充”，修改颜色为白色。

⑥ 适当调整图例的位置、形状，调整绘图区的大小，调整整个图表区域的字体大小，去掉图表区边框等，可获得如图 6-18 右侧所示的百分条图。

6.6.4　饼图

饼图（pie chart），有时也叫圆图。和百分条图一样，饼图、圆环图也是用来描述百分比构成情况的一种统计图。比如对于例 6-4 可分别做出男、女的饼图。下面给出男性胃癌患者病理组织学类型构成比的饼图绘制方法。

① 选取单元格区域 B3:D4，单击“插入→饼图→二维饼图→饼图（饼图中第 1 个子类型）”。

② 慢速单击某一数据点标志（扇形）两次（第一次是所有数据点），单击鼠标右键，在快捷菜单中选择“设置数据点格式→填充→图案填充”，适当选择前景色、背景色与图

案；单击左侧“边框颜色”，选择边框颜色为“实线”，颜色为“黑色”，单击“确定”；继续以同样方式选取另外两个数据点（扇形）标志并进行类似修改。

③在任意一数据（扇形）点标志处单击左键，整个圆被选取，此时单击选项卡“图表工具→布局→标签→数据标签→其他数据标签选项”，弹出“设置数据标签格式”对话框，在标签选项右侧的“标签包括”下勾选“百分比”和“显示引导线”，“标签位置”下勾选“居中”；单击左侧“数字”，在对话框右侧“类别”下选择“数字”，并将小数位数改为 1。单击左侧“填充”，选择“纯色填充”，将数值标签区域的颜色选为白色。

④适当调整图例的位置、形状，调整绘图区的大小，调整整个图表区域的字体大小等可获得如图 6-19 所示的效果。

⑤复制图 6-19，粘贴图形到该工作表的某一位置，在新复制的图形图表区单击鼠标右键，选择快捷菜单中“更改图表类型”的“圆环图→分离型圆环图（圆环图第 2 个子类型）”，单击“确定”，适当调整修饰图形，可获得如图 6-20（也见光盘文件“06-07 饼图.xlsx”）所示的效果。

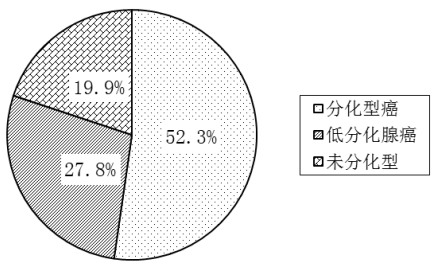


图 6-19　男性胃癌分类饼图

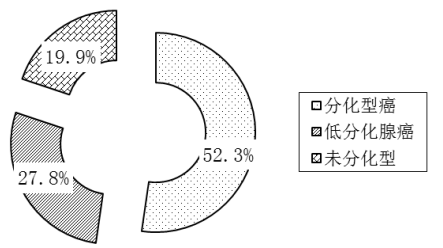


图 6-20　男性胃癌分类分离形圆环图

6.6.5　复合条饼图

有时除了对某种属性的构成感兴趣外，还对该属性中某一类的构成感兴趣，此时需要绘制复合条饼图（bar of pie chart）。

例 6-5：卫生机构包括医院、卫生院、疾控中心、妇幼保健院和其他，其中每个机构人员包括卫生技术人员、其他技术人员、管理人员、工勤人员，现在除对卫生机构构成感兴趣外，还对医院卫生人员构成感兴趣，拟绘制成图 6-21（见光盘文件“06-08　复合条饼图.xlsx”），问应如何绘制？

解：图 6-21 左侧的数据有 8 个分类，其中后面 4 个分类（所在单元格区域为 A7:B10）均属于“医院”。由于医院人数在所有卫生人员中占一半以上，因此也希望在图中具体显示“医院”内部的人员构成（见图 6-21 右侧的百分条图）。该图形的绘制方法如下：

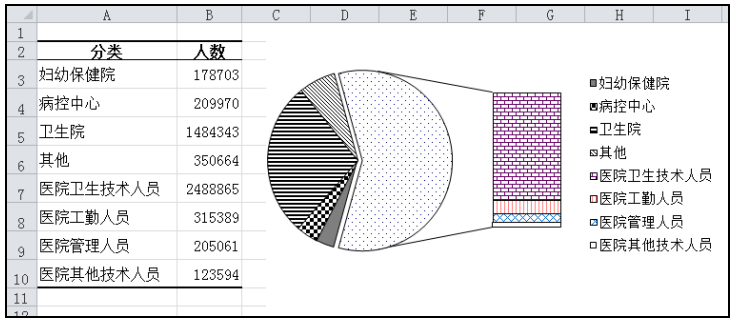


图 6-21　卫生机构及医院卫生人员构成

①选取单元格区域 A3:B10，单击“插入→饼图→二维饼图→复合条饼图（二维饼图中的最后一个子图）”。

②此时 Excel 自动将最后 3 行数据作为条图的数据。显然这是错误的，本例是希望条图显示最后 4 行数据（即医院卫生人员）。为了纠正错误，需要右击饼图，在快捷菜单中选择“设置数据系列格式”，在弹出的对话框中选择“系列选项”，修改“第 2 绘图区包括最后一个”之后的“3”为“4”，单击“关闭”按钮。

③右键鼠标单击每一个数据点（扇形或矩形块），选择“修改数据点格式”，可对每个数据点图案进行修改，以便在图形为黑白灰度情况下能够分辨出来。图表区边框选择“无”。

④为了在饼图中使“医院”部分与其他部分有一定的区分，可以先选取“医院”数据系列，在鼠标显示十字箭头情况下，略向右拖动“医院”扇区。最终获得图 6-21 右侧图形。

⑤复制图 6-21 右侧图形，粘贴该图形到该工作表的右侧，在新复制的图形图表区单击鼠标右键，选择快捷菜单中“更改图表类型→饼图→复合饼图（pie of pie）（第 3 个子图）”，单击“确定”按钮，适当调整图例显示位置，并修饰图形，可获得如图 6-22（也见光盘文件“06-08 复合条饼图.xlsx）所示的效果。

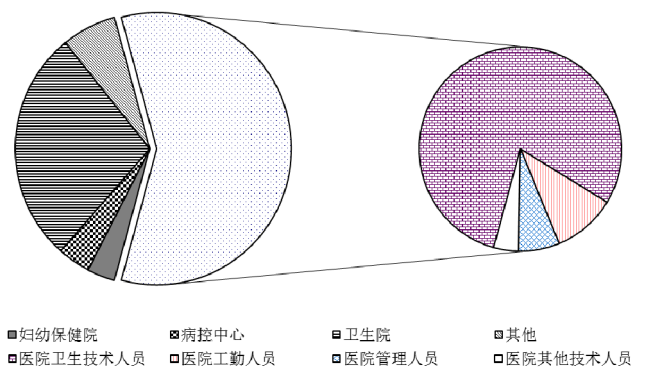


图 6-22　复合饼图

注意：复合饼图的绘制方法与绘制饼图类似。绘制此图最重要的就是将次级构成（如绘制条图）数据放在主要构成数据之后，然后在饼图“设置数据系列格式”中选择“系列选项”，正确修改数据分割位置即可。

6.6.6　线图与半对数线图

线图（line chart）是用点和点之间的连线表示统计指标的变化趋势（绝对差）的统计图，其纵、横轴均为算术尺度。半对数线图（semi-logarithmic linear chart）是线图的一种特殊形式，适用于表示事物发展速度（相对比），其纵轴为对数尺度，横轴为算术尺度。

例 6-6：某市 1991～2001 年婴儿死亡率与孕产妇死亡率见图 6-23（见光盘文件“06-09 线图与半对数线图.xlsx”）中的表格，请绘制线图与半对数线图，分别反映两个死亡率指标的变化趋势与变化速度。

	A	B	C
1	表 某市1991～2001年婴儿与孕产妇死亡率（‰）		
2		死亡率（‰）	
3	年份	婴儿	孕产妇
4	1991	14.27	0.32
5	1992	18.83	0.37
6	1993	16.23	0.34
7	1994	14.97	0.30
8	1995	16.65	0.30
9	1996	16.18	0.31
10	1997	16.15	0.27
11	1998	13.98	0.26
12	1999	11.82	0.29
13	2000	10.88	0.18
14	2001	8.15	0.23
15			

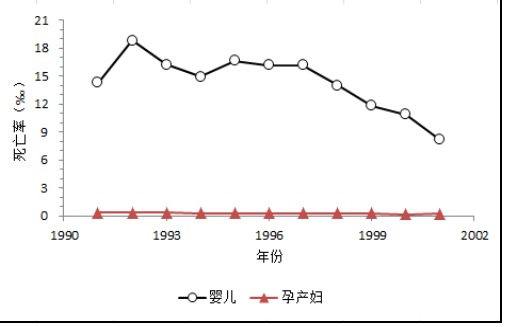


图 6-23　某市 1991～2001 年婴儿与孕产妇死亡率（‰）

解：按下列操作完成线图与半对数线图的制作。

① 按图 6-23 键入单元格区域 A1:C14 中文字与数据。选取单元格区域 A3:C14，单击“插入→散点图→带直线和数据标记的散点图（散点图中第 4 个子图）”。

② 在网格线上单击鼠标右键，选择“删除”。

③ 在婴儿的数据系列（线）标志上单击鼠标右键，选择“设置数据系列格式→数据标记选项”，将“数据标记类型”由“自动”改为“内置”，在“类型”下拉菜单中选择实心圆图形；单击“数据标记填充→纯色填充”，将填充颜色修改为白色。修改“线条颜色”、“标记线颜色”均为黑色。在此对话框还可更改前景颜色、样式的大小、线条的形状粗细等，单击“关闭”。如果需要改变另一个数据系列格式，可仿此设置。

④ 在横坐标上单击鼠标右键，选择“设置坐标轴格式→坐标轴选项”，设置最小值、

最大值、主要刻度单位、次要刻度单位分别为 1990、2002、3、0.6；“主要刻度线类型”与“次要刻度线类型”均选中“外部”；选择“纵坐标轴交叉→坐标轴值”，填写“1990”。在纵轴及其标题区内单击鼠标右键，选择“设置坐标轴格式→坐标轴选项”，设置最小值、最大值、主要刻度单位、次要刻度单位分别为 0、21、3、1；“主要刻度线类型”与“次要刻度线类型”均选中“外部”；选择“纵坐标轴交叉→坐标轴值”，填写 0；选择“数字”，在“类别”下拉列表中选择“数字”，小数位数后改变为“0”，负数选择“–1234”。单击“关闭”。

⑤适当调整图例的位置、形状，添加横、纵坐标轴标题及修改字体，调整绘图区的大小，调整整个图表区域的字体大小等，去掉图表区外面的边框，可获得图 6-23 右侧的线图。

⑥复制图 6-23 右侧图形，粘贴图形到该工作表的某一位置，在新复制的图形的纵坐标轴及其标题区内单击鼠标右键，选择“设置坐标轴格式→坐标轴选项”，设置最小值、最大值、主要刻度单位、次要刻度单位分别为 0.1、100、10、1；单击“对数刻度”前的小方框；选择“纵坐标轴交叉→坐标轴值”，填写 0.1。单击“确定”按钮，可获得如图 6-24 所示的效果。

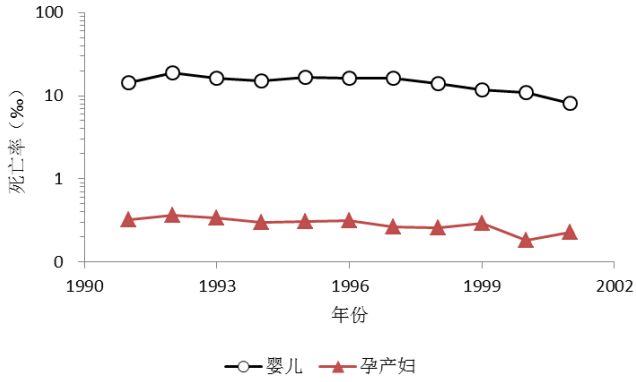


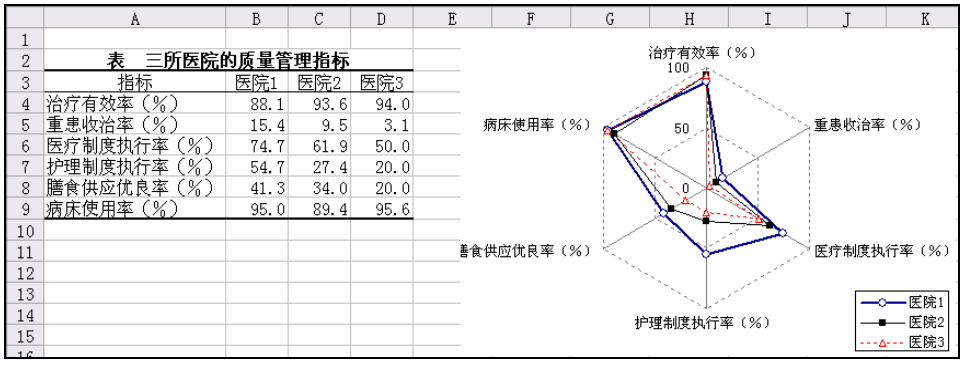
图 6-24　半对数线图

结果分析：将同一资料分别绘制成普通线图（见图 6-23）与半对数线图（见图 6-24），呈现两种不同的结果。图 6-23 中两条线的坡度相差悬殊，是由于婴儿死亡率与孕产妇死亡率绝对值相差悬殊之故，反映了婴儿死亡率与孕产妇死亡率各自的变化趋势；而图 6-24 中两条线的坡度基本一致，是由于婴儿死亡率与孕产妇死亡率的对数值相差（即相对比）相等，反映了婴儿死亡率与孕产妇死亡率的变化速度基本相同。

6.6.7　雷达图

雷达图（radar graph）可比较多个单位在不同指标之间的数据特征，如最大值、最小值等。

例 6-7：为了反映医院管理的质量，某市调查了几所医院的 6 项指标（指标值越大表示质量越高），其中三所医院的调查结果见图 6-25（也见光盘文件“06-10 雷达图.xlsx”）左侧表格，请绘制雷达图反映三所医院的管理质量。



	A	B	C	D
1				
2	表　三所医院的质量管理指标			
3	指标	医院1	医院2	医院3
4	治疗有效率（%）	88.1	93.6	94.0
5	重患收治率（%）	15.4	9.5	3.1
6	医疗制度执行率（%）	74.7	61.9	50.0
7	护理制度执行率（%）	54.7	27.4	20.0
8	膳食供应优良率（%）	41.3	34.0	20.0
9	病床使用率（%）	95.0	89.4	95.6
10				
11				
12				
13				
14				
15				

图 6-25　医院管理质量雷达图

解：绘制方法如下：

① 按图 6-25 键入单元格区域 A2:D9 中文字与数据，略加编辑修饰。选取单元格区域 A3:D9，单击“插入→其他图表→雷达图→带数据标记的雷达图（雷达图的第 2 个子图）”。图形从中心向四周共射出 6 条数值坐标轴，分别代表治疗有效率、重患收治率等 6 个指标。图中的等边六边形是 6 个数值坐标轴的主要网格线。

② 在数值坐标轴的主要网格线上单击鼠标右键，选择“设置主要网格线格式→线条颜色”在对话框右侧选中“实线”，将颜色变为自己想要的颜色，如灰色（白色，背景 1，深色 50%）。单击对话框左侧“线型”，将“宽度”设为 0.25 磅，“复合类型”设为“单线（默认）”，“短划线类型”选为方点（第 3 条），单击“关闭”按钮。

③ 在坐标轴任一数值标签上单击鼠标右键，选择“设置坐标轴格式→坐标轴选项”，可按自己的意愿设置最小值、最大值、主要刻度单位、次要刻度单位等，本例分别设为 0、100、50、50；选择“数字”，在“类别”下拉列表中选择“数字”，小数位数后改变为“0”，负数选择“–1234”；单击对话框左侧“线条颜色”，将颜色变为自己想要的颜色，如灰色（白色，背景 1，深色 50%），单击“关闭”按钮。

④ 在医院 1 数据系列（点或线）标志上单击鼠标右键，选择“设置数据系列格式→数据标记选项”，在“数据标记类型”选择“内置”，类型选择圆圈（第 8 个），“大小”设置为 6；单击左侧“数据标记填充”，选择“纯色填充”，颜色设置为“白色”；单击左侧“线条颜色”，选择“实线”，颜色选择深蓝；单击左侧“线型”，将“宽度”设为 2 磅，“复合类型”设为“单线（默认）”，“短划线类型”选为实线（第 1 条），单击“关闭”按钮。其他两个医院的数据系列格式可仿此设置。

⑤适当调整图例的位置、形状，调整绘图区的大小，调整整个图表区域的字体大小等，可获得图 6-25 右侧的雷达图。

结果分析：从图 6-25 可见，医院 1 的 6 个指标大小的大部分都超过医院 2 和医院 3，而医院 2 的大部分又超过医院 3。这就是说：此次抽样调查获得的医院管理质量从高到低排序分别为医院 1、医院 2、医院 3。

6.6.8 气泡图

气泡图（bubble diagram）最多可以表示五维（x 位置、y 位置、大小、颜色和时间），通过更改气泡的大小和颜色，按时间变化将气泡制成动画视觉效果，能使数据探索更加方便。

例 6-8：某地采用直接测量方法，测得年龄在 0～15 周岁的不同性别儿童的身高、体重、体表面积等指标，从中随机挑选的 18 名儿童的测量值如图 6-26（也见光盘文件“06-11 气泡图.xlsx”）所示。试对该数据采用气泡图反映身高、体重、体表面积 3 个指标间的关系。

	A	B	C
1	体重(kg)	身高(cm)	表面积(m^2)
2	3	51	0.1850
3	8	77	0.3934
4	15	94	0.6075
5	16	105	0.6785
6	20	114	0.7433
7	25	130	0.9028
8	32	140	1.0584
9	32	144	1.0822
10	41	153	1.2627
11	2	51	0.1928
12	10	76	0.3846
13	16	99	0.6411
14	16	105	0.6407
15	25	123	0.8756
16	25	124	0.8855
17	30	141	1.0291
18	36	143	1.1282
19	45	157	1.3427
20			

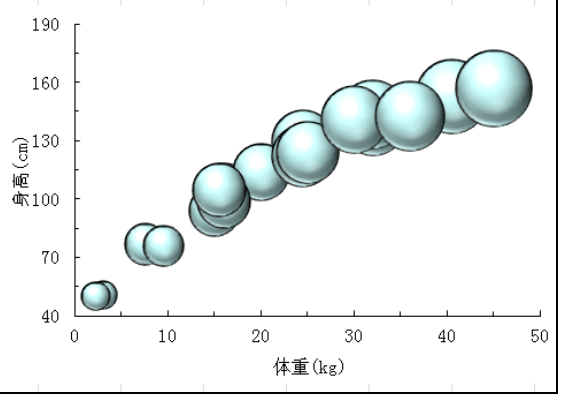


图 6-26　身高、体重与人体表面积的关系

解：绘制方法如下：

①按图 6-26 键入单元格区域 A1:C19 中的文字与数据，略加编辑修饰。取单元格区域 A2：C19，单击“插入→其他图表→气泡图→三维气泡图（气泡图的第 2 个子图表类型）”。

②在网格线上单击鼠标右键，选择“删除”；在图例上单击鼠标右键，选择“删除”。

③单击功能表“图表工具→布局→标签→坐标轴标题→主要横坐标轴标题→坐标轴下方标题”，在数值（X）轴下方填入“体重（kg）”；同样单击“主要纵坐标轴标题→旋转过的标题”在数值（Y）轴左侧填入标题“身高（cm）”。

④在数据系列（气泡）标志上单击鼠标右键，选择“设置数据系列格式”，在弹出的对话框中设置相应的填充颜色和边框。单击“关闭”按钮。

⑤在横坐标轴及其标题区内单击鼠标右键，选择“设置坐标轴格式→坐标轴选项”，设置最小值、最大值、主要刻度单位、次要刻度单位分别为 0、50、10、5，“主要刻度线类型”与“次要刻度线类型”均选择“内部”；选择“纵坐标轴交叉→坐标轴值”，填写 0，单击“关闭”按钮。在纵坐标上单击鼠标右键，选择“设置坐标轴格式→坐标轴选项”，设置最小值、最大值、主要刻度单位、次要刻度单位分别为 40、190、30、15，“主要刻度线类型”与“次要刻度线类型”均选择“内部”；选择“横坐标轴交叉→坐标轴值”，填写 40，单击“关闭”按钮。

⑥适当调整绘图区的大小，调整整个图表区域的字体大小等，可获得图 6-26 右侧的气泡图。

结果分析：由图 6-26 可见，在 0～15 岁儿童中，随着体重的增大，身高逐渐增加，但后期增加幅度有所减小；随着体重、身高的逐渐增大，气泡逐渐增大，即体表面积逐渐增大。

例 6-9：某研究者养了 9 只大白鼠，5 只鼠未服药，2 只鼠用泰舒达治疗，2 只鼠用美多巴治疗，资料见图 6-27 左侧（也见光盘文件“06-12 气泡图 2.xlsx”）。试采用气泡图反映不同治疗方法在不同时间、不同部位的左比值变化情况。

	A	B	C	D	E
1	编号	治疗	时间（周）	部位	左比值
2	1	未服药	0	1	0.1206
3	2	未服药	0	2	0.1034
4	3	未服药	4	1	0.1014
5	4	未服药	4	2	0.0995
6	5	未服药	8.5	1	0.1710
7	6	泰舒达	9.5	1	0.1068
8	7	泰舒达	9.5	2	0.0270
9	8	美多巴	9	1	0.0487
10	9	美多巴	8.5	2	0.0377
11					

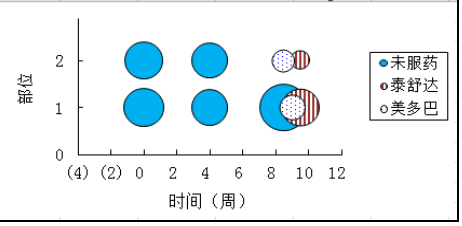


图 6-27　3 种治疗方案不同时间、不同部位的左比值变化情况

解：绘制方法如下：

①单击“插入→其他图表▼→气泡图→气泡图（气泡图的第 1 个子图）”。

②在图表区或绘图区单击鼠标右键，在快捷菜单中单击“选择数据”，弹出“选择数据源对话框”，单击“图例项（系列）”下的“添加”，在“系列名称”右侧用鼠标选取单元格 B2，在“X 轴系列值”右侧用鼠标选取单元格区域 C2:C6，在“Y 轴系列值”右侧用鼠标选取单元格区域 D2:D6，在“系列气泡大小”右侧用鼠标选取单元格区域 E2:E6，单击“确定”、“添加”，在“系列名称”右侧用鼠标选取单元格 B7，在“X 轴系列值”右侧用鼠标选取单元格区域 C7:C8，在“Y 轴系列值”右侧用鼠标选取单元格区域 D7:D8，在“系列气泡大小”右侧用鼠标选取单元格区域 E7:E8；再单击“添加”，在“系列名称”右侧用鼠标选取单元格 B9，在“X 轴系列值”右侧用鼠标选取单元格区域 C9:C10，在“Y

轴系列值”右侧用鼠标选取单元格区域 D9:D10，在“系列气泡大小”右侧用鼠标选取单元格区域 E9:E10。

③单击“图表工具→布局→坐标轴标题→主要横坐标轴标题→坐标轴下方标题”，在数值（X）轴下方填入“时间（周）”；同样单击“主要纵坐标轴标题→旋转过的标题”在数值（Y）轴左侧填入标题“部位”。

④在网格线上单击鼠标右键，选择“删除”。

⑤在横坐标轴及其标题区内单击鼠标右键，选择“设置坐标轴格式→坐标轴选项”，设置最小值、最大值、主要刻度单位、次要刻度单位分别为–4、12、2、2，“主要刻度线类型”与“次要刻度线类型”均选择“内部”；选择“纵坐标轴交叉→坐标轴值”，填写–4；选择“数字”，在“类别”下拉列表中选择“数字”，小数位数后改变为“0”，负数选择“（1234）”，单击“关闭”按钮。

⑥在纵坐标轴及其标题区内单击鼠标右键，选择“设置坐标轴格式→坐标轴选项”，设置最小值、最大值、主要刻度单位、次要刻度单位分别为 0、2.9、1、1，选择“横坐标轴交叉→坐标轴值”，填写 0，单击“关闭”按钮。

⑦单击“未服药”系列标志，选择“图表工具→格式→形状填充”，选择浅蓝色。右键鼠标单击“泰舒达”系列标志，选择“设置数据系列格式→填充→图案填充”，在前景颜色处选择深红色（红色，强调文字颜色 2，深色 25%），背景颜色处选择白色，选择第 5 排第 4 个图案，单击“关闭”；在“美多巴”系列标志上单击鼠标右键，选择“设置数据系列格式→填充效果→图案填充”，在前景颜色处选择深蓝色（蓝色，强调文字颜色 1，深色 25%），背景颜色处选择白色，选择第 2 排第 1 个图案，单击“关闭”按钮。

⑧适当调整绘图区的大小，调整整个图表区域的字体大小，去除图表区的边框等，可获得图 6-27 右侧的气泡图。

结果分析：由图 6-27 可见，美多巴治疗的两个部位、泰舒达治疗的部位 2，其左比值气泡变小，即它们的左比值数量有所减少。

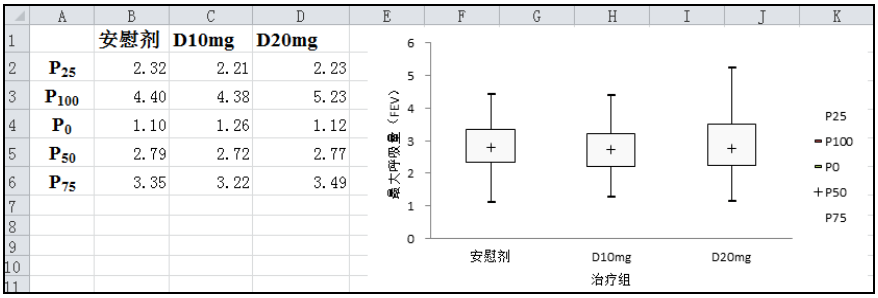
6.6.9 箱图

箱图（boxplot）也称箱须图（box-whisker plot），用于反映一组或多组连续型定量数据分布的集中位置和离中散布范围。箱图的中心位置为中位数（第 50%百分位数）；中部的“箱”范围为四分位间距（即 P_{75}–P_{25}）；“箱”两端的“须”一般为最大值与最小值，如果资料两端值变化较大，两端也可采用 $P_{99.5}$ 与 $P_{0.5}$、P_{99} 与 P_1、$P_{97.5}$ 与 $P_{2.5}$。读者可根据数据的波动情况作出选择。这里的 P 表示百分数符号，如“P_{75}”表示 75%百分位数值。

（1）用股价图制作箱图

例 6-10：某研究者分别采用安慰剂、新药 10mg、新药 20mg 治疗三个随机分组的

阻塞性肺病病人，每组 100 例，治疗两周后测量最大呼气量（Forced Expiratory Volume, FEV）。每组测量后计算获得的最大值（P_{100}）、75%百分位数（P_{75}）、中位数（P_{50}）、25%百分位数（P_{25}）、最小值（P_0）见图 6-28（也见光盘文件“06-13 箱图.xlsx”）所示，试绘制该箱图。

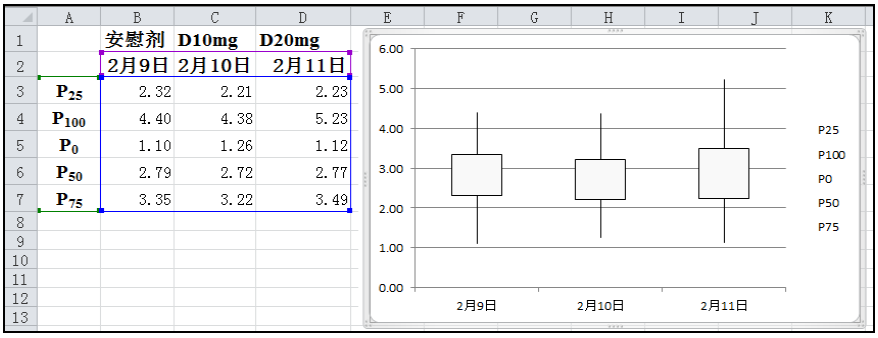


	A	B	C	D
1		安慰剂	D10mg	D20mg
2	P_{25}	2.32	2.21	2.23
3	P_{100}	4.40	4.38	5.23
4	P_0	1.10	1.26	1.12
5	P_{50}	2.79	2.72	2.77
6	P_{75}	3.35	3.22	3.49

图 6-28　三种方案治疗阻塞性肺病后的最大呼气量

解：绘制箱图需要借助于股价图中的“开盘—盘高—盘低—收盘图”，该图形需要将数据按一定的顺序排列。绘制箱图时需要将数据按 P_{25}、P_{100}、P_0、P_{75}、P_{50} 的顺序排列。绘制方法如下：

① 在第一行下方插入一行，在 B3 至 D3 随意输入三个不同的日期（插入股价图所要求的格式，图表制作完成后可删除），如图 6-29 所示。选取单元格区域 B2:D6，单击“插入→其他图表→股价图→开盘—盘高—盘低—收盘图（股价图的第 2 个子图）”。



	A	B	C	D
1		安慰剂	D10mg	D20mg
2		2月9日	2月10日	2月11日
3	P_{25}	2.32	2.21	2.23
4	P_{100}	4.40	4.38	5.23
5	P_0	1.10	1.26	1.12
6	P_{50}	2.79	2.72	2.77
7	P_{75}	3.35	3.22	3.49

图 6-29　箱图的初始界面

② 在“选择数据源”对话框中单击“添加”，在“系列名称”下用鼠标选取单元格 A7，在“值”右侧用鼠标选取单元格区域 B7:D7，单击“确定”按钮。右键单击图表区或绘图区，再选择“选择数据”，在“选择数据源”对话框中单击“系列 1→编辑”，在“系列名称”下用鼠标选取单元格 A3，单击“确定”按钮。再用同样的方式修改系列 2 至系列 4 的名称分别为 A4、A5、A6。获得图 6-29 的界面。

在第 2 行上单击鼠标右键，选择“删除→删除整行”，去掉刚才加入的第 2 行“日期”数据（注意，删除时 Excel 弹出提示存在多个无效链接情况，可忽略这些提示）。

③单击“水平（分类）轴标签”下“编辑”，在弹出的对话框中“轴标签区域”下方用鼠标选取单元格区域 B1:D1，单击“确定”按钮。在“选择数据源”对话框中单击“确定”按钮关闭对话框。

④选择“图表工具→布局→坐标轴标题主要横坐标轴标题→坐标轴下方标题”，设置分类（*X*）轴下方标题“治疗组”，同样单击“主要纵坐标轴标题→旋转过的标题”设置数值（*Y*）轴标题为“最大呼气量（FEV）”。

⑤在网格线上单击鼠标右键，选择“删除”。

⑥在横坐标轴及其标题区内单击鼠标右键，选择“设置坐标轴格式→坐标轴选项”，“主要刻度线类型”选择“无”，单击“确定”按钮。

⑦在纵坐标轴及其标题区内单击鼠标右键，选择“设置坐标轴格式→数字→类别→数字”，小数位数改为“0”，单击“关闭”按钮。

⑧在鼠标移至箱的中心位置，系列标志显示为 P_{50} 时单击鼠标右键，选择“设置数据系列格式”，在弹出的对话框左侧选择“数据标记选项”标签，在界面右侧数据标记类型处选“内置”，并选择“+”，“大小”改为 6 磅，单击左侧“标记线颜色”，在右侧选择“实线”，颜色处选择黑色，单击“关闭”按钮。

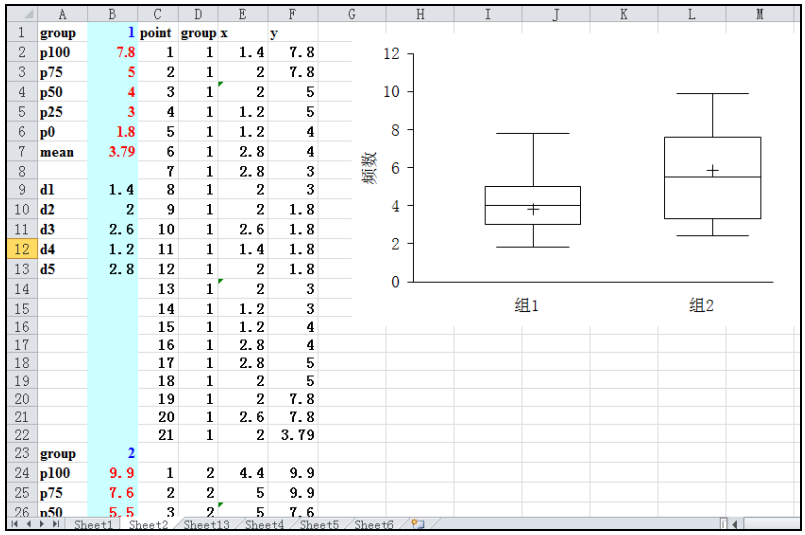
在箱图的顶部系列标志（P_{100}）上单击鼠标右键，选择“设置数据系列格式→数据标记选项”，在界面右侧数据标记类型处选“内置”以及长横线“－”；单击左侧“标记线颜色”，在右侧选择“实线”，颜色处选择黑色，单击“关闭”。在箱图的底部系列标志（P_0）上单击鼠标右键，在弹出的对话框左侧选择“数据标记选项”标签，在界面右侧数据标记类型处选“内置”以及长横线“－”；单击左侧“标记线颜色”，在右侧选择“实线”，颜色处选择黑色，单击“关闭”按钮。

⑨适当调整绘图区的大小，调整整个图表区域的字体大小，去除图表区的边框等，可获得图 6-28 右侧的箱图。

（2）用散点图制作箱图

光盘文件“06-14 箱图 2.xlsx”中共有 6 个工作表，每个工作表中分别包含了 1～6 个箱图。现以在一个统计图中制作两个箱图为例，来说明“箱图”的制作方法。

假定某个箱图的 P_{100}、P_{75}、P_{50}、P_{25}、P_0 以及均数（mean）已知，第 1 组（group）的这些值位于单元格区域 B2:B7。为了绘制一个箱图需要有 21 个数据点，数据点坐标的 *X* 值（比较 B9:B13 与 E2:E22）控制箱图的宽窄（见单元格 E2:E22），*Y* 值显示箱图的 P_{100}、P_{75}、P_{50}、P_{25}、P_0 以及均数值（见单元格 F2:F22）。E2:E21 与 F2:F21 为系列 1（采用散点图的第 5 个子图）；E22 与 F22 为系列 2（采用散点图的第 1 子图），将数据标志样式改为“＋”，大小改为 8 磅。



	A	B	C	D	E	F
1	group	1	point	group	x	y
2	p100	7.8	1	1	1.4	7.8
3	p75	5	2	1	2	7.8
4	p50	4	3	1	2	5
5	p25	3	4	1	1.2	5
6	p0	1.8	5	1	1.2	4
7	mean	3.79	6	1	2.8	4
8			7	1	2.8	3
9	d1	1.4	8	1	2	3
10	d2	2	9	1	2	1.8
11	d3	2.6	10	1	2.6	1.8
12	d4	1.2	11	1	1.4	1.8
13	d5	2.8	12	1	2	1.8
14			13	1	2	3
15			14	1	1.2	3
16			15	1	1.2	4
17			16	1	2.8	4
18			17	1	2.8	5
19			18	1	2	5
20			19	1	2	7.8
21			20	1	2.6	7.8
22			21	1	2	3.79
23	group	2				
24	p100	9.9	1	2	4.4	9.9
25	p75	7.6	2	2	5	9.9
26	p50	5.5	3	2	5	7.6

图 6-30　两个箱图

系列 1 的 E2:E11 与 F2:F11 绘制的线图（见图 6-31 左侧）和 E12:E21 与 F12:F21 绘制的线图（见图 6-31 右侧）重叠恰好制成第一个箱图的形状。

第 2 个箱图可仿照第 1 个箱图进行制作，但应特别注意：数据点坐标的 X 值在第 1 个图基础上加了“3”。如果制作第 3 个箱图，则在第 2 个图基础上加“3”，依此类推。

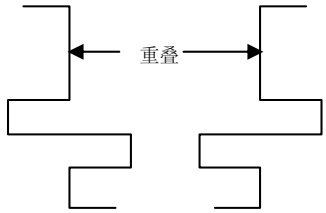


图 6-31　箱图的绘制原理

根据需要制作箱图的多少，可选择不同的工作表。通过修改第 B 列的 P_{100}、P_{75}、P_{50}、P_{25}、P_0 以及均数（mean）数据，可以方便地制作你自己的箱图。

6.6.10　人口金字塔图

人口金字塔（population pyramid）图是利用人口的性别、年龄分组数据绘制的人口构成图，通常年龄为纵轴，人数或其百分数为横轴，纵轴左侧为男，右侧为女。

例 6-11：表 6-2 为我国 1964 年与 2000 年的人口性别与年龄构成人数，相应的金字

塔见图 6-32a、图 6-32b（也见光盘文件“06-15 金字塔.xlsx”）。用 Excel 绘制这两个图形。

表 6-2　中国 1964 年与 2000 年的年龄、性别人口数（万）

年龄（岁）	1964 年		2000 年	
	男	女	男	女
0～	3765	3133	3765	3133
5～	4830	4185	4830	4185
10～	6535	6005	6535	6005
15～	5288	5015	5288	5015
20～	4794	4664	4794	4664
25～	6023	5737	6023	5737
30～	6536	6195	6536	6195
35～	5614	5301	5614	5301
40～	4224	3900	4224	3900
45～	4394	4158	4394	4158
50～	3280	3050	3280	3050
55～	2406	2231	2406	2231
60～	2168	2003	2168	2003
65～	1755	1723	1755	1723
70～	1244	1314	1244	1314
75～	718	875	718	875
80～	320	479	320	479
85～	134	266	134	266

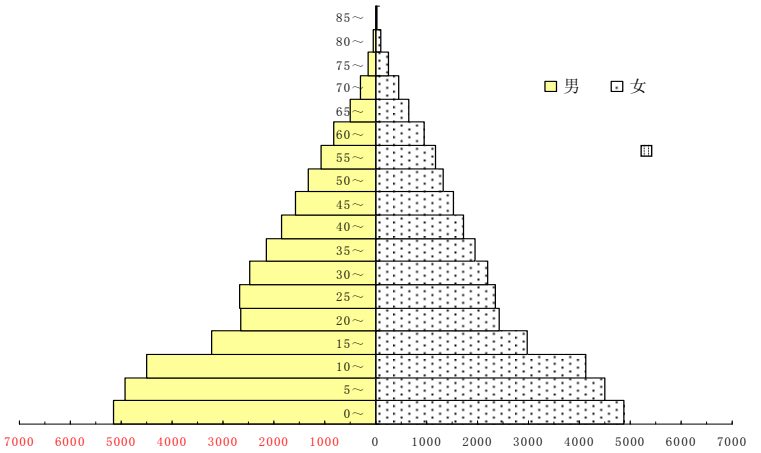


图 6-32a　1964 年人口金字塔

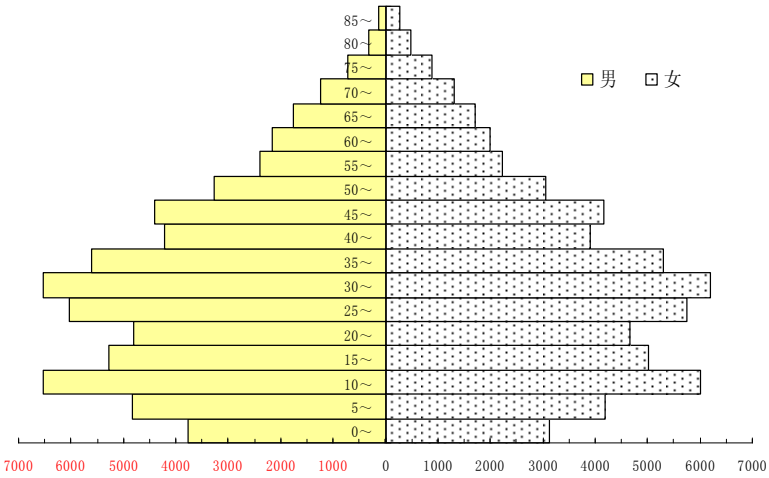


图 6-32b　2000 年人口金字塔

解：下面以图 6-32b 为例说明金字塔图绘制的方法。绘制步骤如下：

①数据放在光盘文件“06-15 金字塔.xlsx”的“2000 年”工作表中，单元格区域 A2:C20 的 A、B、C 列分别存放年龄、男性人口数、女性人口数。

②为了用 Excel“条形图”平行地将男、女数据绘制在中心轴两侧，首先必须将放在左侧的数据（男性人口数）变成负数，且不显示正负符号。本例在 F1 单元格存放“-1”，复制 F1，选取男性人口数单元格区域 B3:B20，在“开始→粘贴→选择性粘贴”，在弹出的对话框中选择“运算→乘”，单击“确定”。再次选取男性人口数单元格区域 B3:B20，右击后选择快捷菜单中的“设置单元格格式→数字→数值→负号（1234）（第 3 个，显示为红色）”，小数位数设置为 0。

③选取单元格区域 A2:C20，单击“插入→条形图→簇状条形图（条形图第 1 个子图）”。

④在网格线上单击鼠标右键，选择“删除”。

⑤在水平横坐标轴及其标题区内单击鼠标右键，选择“设置坐标轴格式→坐标轴选项”，设置最小值、最大值、主要刻度单位、次要刻度单位分别为-7000、7000、1000、500；“次要刻度线类型”选择“内部”；在“纵坐标交叉→坐标轴值”之后填写 0；选择“数字→类别→数字”，小数位数设置为 0，单击“关闭”按钮。适当放大图形，图表区字体选择 9 号。

⑥在垂直中心线轴及其标题区内单击鼠标右键，选择“设置坐标轴格式→坐标轴选项”，主、次要刻度线类型均选择“无”。选择“刻度线间隔”，“标签间隔”、指定间隔单位”、“标签与坐标轴的距离”，均设置为“1”。单击“关闭”按钮。

⑦在男性人口数条图上单击鼠标右键，选择“设置数据系列格式→系列选项”，“系列重叠”比例为“100”，设置数据系列格式的“填充→纯色填充→颜色”为黄色，“边框

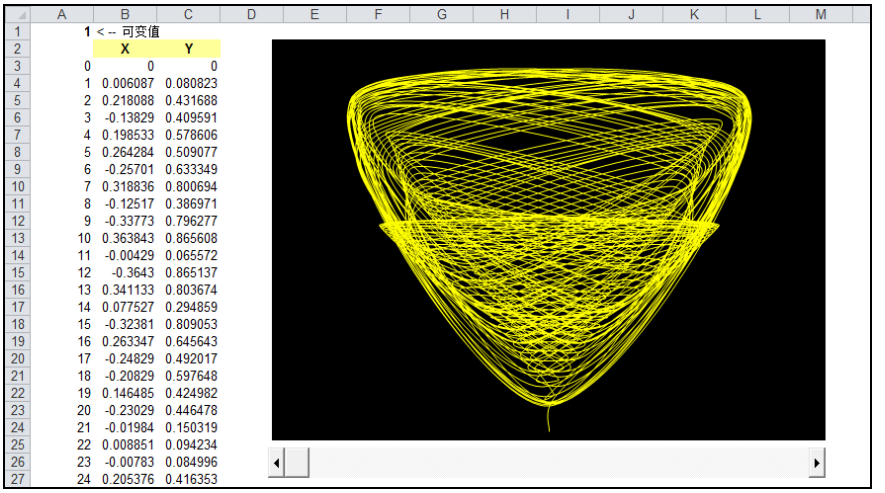
颜色”设置为黑色，分类间距为“0”。单击“关闭”。

⑧在女性人口数条图上单击鼠标右键，选择“设置数据系列格式→填充→图案填充”，选择第 2 行第 1 列图案，前景色为“黑色”，背景色为“白色”，“边框颜色”设置为黑色，单击“关闭”按钮。

图 6-32b 制作好后，可在“2000 年”工作表标签名称上单击鼠标右键，选择“移动或复制”，在对话框左下角勾选“建立副本”，单击“确定”按钮，工作表标签名称修改为“1964 年”，将表 6-2 中有关数据替代单元格区域 B2:C20。这样即可制成图 6-32a。

结果分析：人口金字塔的形状反映了长期以来人口的出生、死亡、迁入、迁出等情况。就 1964 年金字塔而言，抗日战争出生的人口集中在 20～24 岁，该组的男性出现了塔形的内凹，反映了抗日战争时期及以后岁月男性死亡较多的情况，另外可以看到 14 岁以下人口增长迅猛，金字塔基底增宽。36 年后，2000 年金字塔有了明显的变化，35～39 岁以上的塔形特征与 1964 年金字塔相似。30～34 岁组人口比例最大，反映了 20 世纪 60 年代末的生育高峰。1973 年 7 月 16 日，国务院成立计划生育领导小组，加大了计划生育的宣传力度，人口出生逐年呈下降趋势。10～15 岁高峰主要由 60 年代出生者（30～34 岁年龄组）达到生育年龄所致。

6.6.11 三角函数制作线图



	A	B	C
1		1 <-- 可变值	
2		X	Y
3	0	0	0
4	1	0.006087	0.080823
5	2	0.218088	0.431688
6	3	-0.13829	0.409591
7	4	0.198533	0.578606
8	5	0.264284	0.509077
9	6	-0.25701	0.633349
10	7	0.318836	0.800694
11	8	-0.12517	0.386971
12	9	-0.33773	0.796277
13	10	0.363843	0.865608
14	11	-0.00429	0.065572
15	12	-0.3643	0.865137
16	13	0.341133	0.803674
17	14	0.077527	0.294859
18	15	-0.32381	0.809053
19	16	0.263347	0.645643
20	17	-0.24829	0.492017
21	18	-0.20829	0.597648
22	19	0.146485	0.424982
23	20	-0.23029	0.446478
24	21	-0.01984	0.150319
25	22	0.008851	0.094234
26	23	-0.00783	0.084996
27	24	0.205376	0.416353

图 6-33　三角函数制作线图

图 6-33（也见光盘文件“06-16 三角函数.xlsx”）的制作要点如下：

①设置一个变数。将 A1 单元格数据从 1～20 进行改变将获得不同形状的图形，因为 A4:C302 的数据均受单元格 A1 数值的影响。为了方便改变单元格 A1 数值，特意制作了“滚

动条”。

② “滚动条”的制作方法是：单击功能区选项卡“开发工具”（默认情况下不会显示这个选项卡。它包含的命令对程序员有用。若要显示该选项卡，请选择“文件”、“选项”，然后选择“自定义功能区”，在其右侧区域选中“开发工具”旁边的复选框，然后单击“确定”按钮），选择“插入→ActiveX 控件→滚动条（ActiveX 控件）”，光标变成实心十字，此时在需要的位置拖放便制成滚动条。为了链接单元格 A1，需在滚动条上单击鼠标右键，选择“属性”（或再功能区单击“设计模式”之后单击滚动条，单击功能区“属性”按钮），将“LargeChange”（步幅）右侧修改为“1”，“LinkedCell”（链接单元格）右侧修改为“A1”，“Max”（最大值）修改为“20”，“Min”（最小值）修改为“1”。单击右上方“X”按钮退出属性对话框。适当调整滚动条的大小、位置等后，单击功能区“设计模式”，使其退出激活状态。

③ 在 B2、C2 单元格分别写上 X、Y，表示即将建立绘图所需的 X、Y 两个变量。在 A3 单元格键入“0”，在 A4 单元格键入公式“=A3+$A$1”，将公式向下拖放至 A302；在 B3 单元格键入公式“=SIN(A3)*COS(A3)*COS(TAN(A3))”，在 C3 单元格键入公式“=SIN(B3)*COS(B3)*SIN(A3+B3)*COS(A3-B3)+SQRT(ABS(B3))”，然后将 B3、C3 单元格公式向下拖放至 C302，所有数据便准备完成。

④ 选取单元格区域 B3:C302，单击“插入→图表→散点图▼→带平滑线的散点图（散点图下的第 3 个子图）

对该图进行适当修饰便可获得图 6-33 右侧的图形。

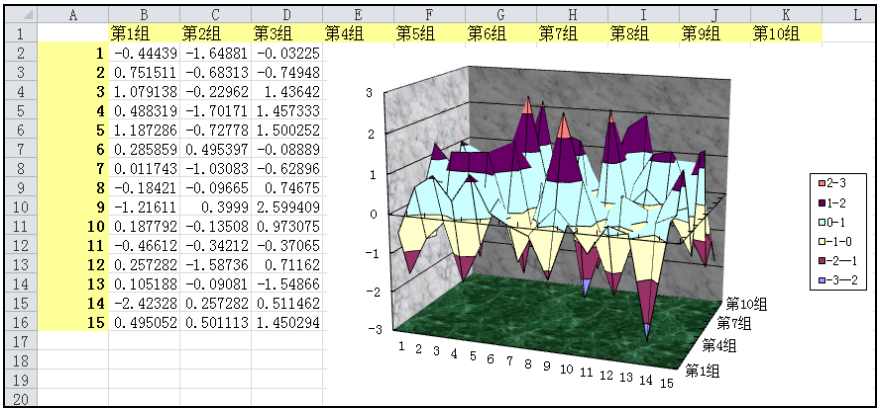
6.6.12　曲面图

例 6-12：用 Excel 随机函数产生均数、标准差分别为 0 与 1 的正态分布数据 10 组，每组 15 个数据，绘制如图 6-34 所示的曲面图（也见光盘文件“06-17 曲面图.xlsx”）。

解：绘制步骤如下：

① 输入图 6-34 的第 A 列和第 1 行。单击功能区选项卡“数据”，选择“数据分析（如果没有需要“加载宏”安装：选择“文件→选项→加载项”，在其右侧区域选择“分析工具库”，单击最下一行中部的“转到”按钮，在弹出的“加载宏”对话框内选择“分析工具库”以及“规划求解加载项”。→随机数发生器”，在弹出的界面中设置变量个数为“10”、随机数个数为“15”、分布选择“正态”、平均值=“0”、标准偏差=“1”、随机数基数可以自己任意设定，不同的随机数基数会产生不同的随机数，输出区域选择“B2”，单击“确定”按钮。

② 选取数据所在单元格区域，单击“插入”其他图表→曲面图→三维曲面图（曲面图的第 1 个子图表）”。



	A	B	C	D
1		第1组	第2组	第3组
2	1	-0.44439	-1.64881	-0.03225
3	2	0.751511	-0.68313	-0.74948
4	3	1.079138	-0.22962	1.43642
5	4	0.488319	-1.70171	1.457333
6	5	1.187286	-0.72778	1.500252
7	6	0.285859	0.495397	-0.08889
8	7	0.011743	-1.03083	-0.62896
9	8	-0.18421	-0.09665	0.74675
10	9	-1.21611	0.3999	2.599409
11	10	0.187792	-0.13508	0.973075
12	11	-0.46612	-0.34212	-0.37065
13	12	0.257282	-1.58736	0.71162
14	13	0.105188	-0.09081	-1.54866
15	14	-2.42328	0.257282	0.511462
16	15	0.495052	0.501113	1.450294
17				
18				
19				
20				

图 6-34　曲面图

③ 设置基底填充效果为“绿色大理石”纹理，背景墙填充效果为“白色大理石”纹理，慢速两次单击每个图例，分别调整图例颜色，曲面图便制成。

从以上绘图过程可见，尽管 Excel 提供了有限的图表标准类型，但通过适当变换、组合，可以产生很多理想图形。

获得 Excel 统计图后，为了复制、粘贴到 Word、PowerPoint 中，通常可采用“开始→粘贴→选择性粘贴”，弹出“选择性粘贴”对话框。有两类选择性粘贴方式：粘贴和粘贴链接。二者各有特点，读者可根据情况，按照对话框下方的“说明”选择进行选择性粘贴。

注意：Excel 的最大特色是图形即改即可见，即数值与图形的变化是联动的，读者可充分利用这一特点，利用以上图形作为模板，修改相应数据，或相应分类标题，便可简单制作自己所需的统计图形。

另外，以上所绘制的图形均没有给出标题，大家在发表文章或撰写论文时，别忘记了统计图的标题。

第 7 章

统 计 描 述

统计学（statistics）是关于数据的设计、搜集、整理、分析和表达的一门学科，目的就是从数据中提取有用信息。总体（population）和样本（sample）是统计学的两个基本概念。根据研究目的所确定的具有相同性质的研究对象的全体就是总体，描述总体的指标称为参数（parameter）；从总体中随机抽取的部分研究对象就是样本(sample)，描述样本的指标称为统计量（statistic）。统计学有描述统计学和推断统计学两大分支，基于样本数据，对总体做出估计、预测或决策的过程就是统计学推断（statistical inference）；采用图形或数据指标等，呈现和小结数据的有关信息的过程就是统计学描述（statistical description）。本章主要介绍区间数据与名义数据的统计学描述方法。

下面先简单介绍如何调用 Excel 的“数据分析”工具，这一工具可帮助我们简单实现统计学的描述与推断。

7.1 数据分析工具的调用

Excel 有一个分析工具叫做“数据分析（data analysis）”，对于统计学工作者及其有关人员十分有用。在安装好 Microsoft Excel 后，一般在 Excel 菜单中找不到这个工具，需要“加载宏”，进行继续安装。

安装“数据分析”工具的方法是：单击“文件→选项”，在“Excel 选项”对话框（见图 7-1）左侧下方单击“加载项”。在右侧下方“管理”下拉列表框中选择“Excel 加载项”。单击“转到”，在弹出的“加载宏”对话框中（见图 7-2）选择“分析工具库”、“分析工具库-VBA”、“规划求解加载项”等，单击“确定”按钮，即可以在 Excel 主界面的“数据”选项卡右侧显示“分析”命令组。

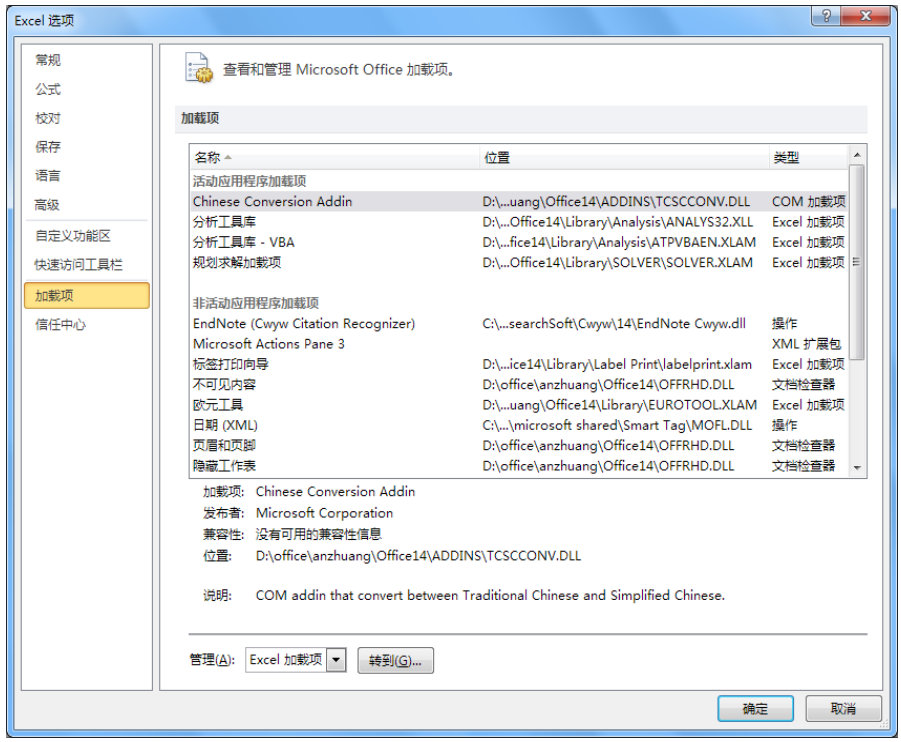


图 7-1　Excel 2010“Excel 选项”对话框

单击“数据→分析→数据分析”，弹出“数据分析”对话框（见图 7-3）。利用“数据分析”工具可简单地完成常用的统计分析，如数据的统计学描述、制作直方图、进行 *t* 检验、方差分析、相关与回归分析等。

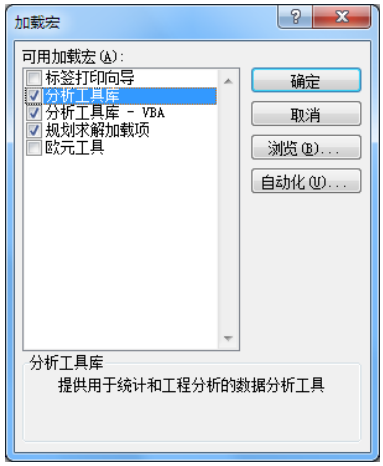


图 7-2　Excel 2010“加载宏”对话框

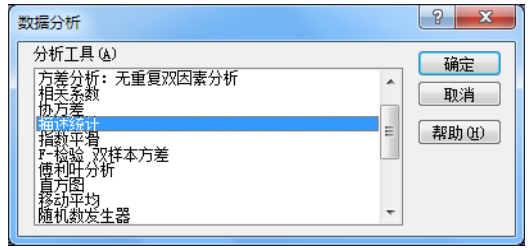


图 7-3　“数据分析”对话框



注意：数据分析工具只对当前的工作表有效，如果要对一个工作簿中的其余工作表进行同样的分析，则应分别在每一张工作表上运行数据分析工具。

7.2　统计学中的数据类型

数据类型、特别是结果变量的数据类型，是数据分析时选择统计学方法的重要依据。在定义数据类型前，需要理解与区分几个基本概念：变量、变量取值和数据。

变量（variable）是总体或样本的一些特征，如儿童的身高是儿童生长发育的一个特征，称身高是儿童生长发育研究的一个变量。并不是每个儿童的身高都相同，不同儿童有不同的身高，所以称之为变量。变量通常是随机变化的，因此也叫随机变量。一个人的性别可能是男性，也可能是女性，这种情况下性别是变量，如果我们研究孕妇的健康状况，此时的性别就不再是变量，而是常量（即性别=女）。

变量取值（values）是变量的可能观察结果，如 6 岁儿童身高的变量取值为 30~200cm（假定儿童的身高最矮为 30cm，最高为 200cm）。

数据（data）是变量的具体观察结果，如观察得到 10 名 6 岁儿童的身高分别为：80、112、105、95、113、100、112、114、102、110（cm）。这些数据就是我们希望从中提取信息的数据。

按测量的精确程度由高到低，数据可分类为：区间（interval）数据、有序（ordinal）数据、名义（nominal）数据三类（见图 7-4）。

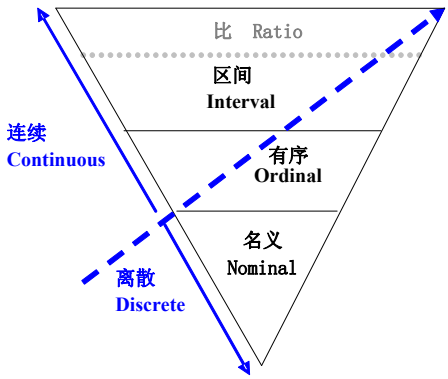


图 7-4　数据的分类

7.2.1　数据类型的辨析

数据的类型直接与统计学方法的选择有关，所以掌握数据的类型具有重要的实际意义。

在区间、有序、名义三类数据中，区间数据是最普通、最常见的数据类别，也叫定量（quantitative）数据或数值（numerical）数据，变量的取值为实数。对于具体的情况，区间数据的取值为一个区间，只要测量工具足够精确，理论上可无限细分。常见的区间数据有：摄氏温度、身高、体重、每分钟心跳次数等。这类数据按严格定义可细分为区间数据和比（ratio）数据两类，摄氏温度就是典型的严格定义的区间变量，尽管两数据间的差值有实际意义，但数值为 0 没有实际意义，0℃为水结冰时的温度，并非绝对意义的零点；身高、体重就是典型的严格定义的比数据，除两数据间差值有实际意义外，数值为 0 也有实际意义。例如，一个人体重为 50kg，另一个人的体重为 100kg，可以说后一个人的体重比前一个人重一倍；昨天气温是 2℃，今天气温是 4℃，就不能说今天比昨天热一倍。统计分析时，将区间数据细分为两类意义不大，所以通常将比数据也归类在区间数据这一类。

名义数据也叫定性（qualitative）数据或分类（categorical）数据，名义变量的变量取值为若干个类别，如 ABO 血型分类为 A、B、O、AB 四类，婚姻状况分类为单身、已婚、离异、独居四类。这些变量取值不是数字，而是文字描述的类别。为了让计算机识别，通常情况下需要对每一类别随意分配一个数字，如将 A、B、O、AB 血型分别编码为 1、2、3、4。只要每一类的数字不同，任何形式的数字编码都是有效的，如对 A、B、O、AB 血型也可编码为-2、-1、1、2 或 3、2、1、0 等。名义数据还可按分类类别数是否为两类，细分为二分类数据和多分类数据。

介于区间与名义二者之间的数据称为有序数据。有序变量与名义变量一样，其变量取值也是若干个类别，但前者的这些类别的取值有大小顺序关系。比如病人病情为轻、中、重；学生成绩为不及格、及格、一般、良好、优秀等。通常情况下需要按类别从小到大的

顺序分配一个数字，比如将学生成绩“不及格、及格、一般、良好、优秀”分别编码为 1、2、3、4、5。编码时必须保持数据的大小顺序关系，在这一条件下的其他编码方法均有效，比如，可将学生成绩“不及格、及格、一般、良好、优秀”分别编码为 30、60、75、85、95。对这些方式编码的数据进行非参数统计分析（见第 9 章），所得的分析结果将完全相同，因为非参数统计学方法考虑的是数据的顺序，而不是数据的具体值大小。

区间数据与有序数据的最大区别在于：任何两个数据观察值之间的差值是否具有实际意义，区间数据之间的差值具有实际意义，而有序数据之间的差值往往不具有实际意义。我们可以合理计算并解释两个区间数据观察值之间的差值，但有序数据则恰好相反，我们不可以合理计算并解释两个有序数据观察值间的差值。上述 1-2-3-4-5 编码方式中，“及格与不及格”之间的差值与“优秀与良好”之间的差值均为 1；而 30-60-75-85-95 编码方式中，“及格与不及格”之间的差值为 30，“优秀与良好”之间的差值为 10。因为这两种编码都是有效的，所以不能采用任何一种编码方式来计算与解释两个有序数据观察值之间的差异。

7.2.2　三类数据的计算

因为名义数据的数字完全是任意给定的，所以不能对名义数据进行直接计算，如 15 名同学的 ABO 血型按 1-2-3-4 编码方式记录的结果分别为：3、2、4、4、3、3、2、2、1、1、3、4、1、2、3，就不能直接对这些数据进行计算。对名义数据进行统计分析，通常的做法是清点每一类别的频数。可以总结以上 15 名同学 ABO 血型数据如表 7-1 所示。

表 7-1　15 名同学ABO血型的频数分布

ABO 血型类别	编　　码	频　　数
A	1	3
B	2	4
O	3	5
AB	4	3

有序数据最重要的特征是变量取值为有序，除了可以制作频数表外，还可以根据数据的顺序关系，确定中间位置的值（即中位数，median）。例如，可由小到大排列数据，选择位置居中的分类类别为平均值。

区间数据允许进行各种计算。除制作频数表、计算中位数外，还可以计算算术平均数、标准差、标准误差等。

7.2.3　三类数据的转化

变量取值较精确的数据可以向变量取值较粗略的数据方向转化，如人的年龄一般为区

间数据，分析时出于实际需要，也可将它转化为有序数据（少年、青年、中年、老年）或名义数据（年轻组、老年组）。反过来不成立，如血型为名义数据，就不能把它当作其他类型数据处理。三类数据的区别如表 7-2 所示。

表 7-2　三类数据的区别

数据类型	变量取值	计算有效性	数据转化
区间	实数	所有计算有效	可转化为有序或名义数据
有序	代表数据大小顺序的数字	基于数据顺序的计算有效（如计算中位数）	可转化为名义数据
名义	代表数据类别的任意数字	基于类别发生频数的计算有效	不能转化

在“第 17 章概率分布”中，将变量分类为离散随机变量和连续随机变量两类。离散随机变量（discrete random variable）的取值可以列举、可以计数。例如抛硬币 10 次，正面朝上的可能取值 x 为 0、1、2、3、4、5、6、7、8、9、10。其值与值之间有缝隙（gap），名义数据一定是离散型随机变量。连续随机变量（continuous random variable）的取值充满整个区间，无法一一列举，如身高、体重等。假定测量很精确，连续随机变量的值与值之间理论上无缝隙。连续变量只能是区间与有序数据，但区间与有序数据也可以是离散随机变量（见图 7-4）。

7.3　区间数据的统计学描述

学生都非常关心考试的成绩，比如考完统计学考试后，考生希望了解全班的考试情况，如 60 分以下、60~69 分、70~79 分、80~89 分、90~100 分的人数各是多少。想获得这一信息，需要学习区间数据的频数表与直方图知识（见 7.3.1 节）；如果希望了解大多数人的考试成绩，需要计算平均数（集中位置指标，见 7.3.2 节）；如果希望知道考试分数是散布在一个较宽的范围（考生成绩相差较悬殊），还是集中在一个较小的范围（考生成绩十分接近），则需要计算标准差（变异度指标，见 7.3.3 节）。

7.3.1　区间数据的频数表与直方图

在同一样本数据中，相同情形出现的次数就是该情形的频数（frequency）。频数在总的观察例数中所占的比例称作相对频数（relative frequency）或频率。若将频数进行累计，可得到累计频数（cumulative frequency）；若将相对频数进行累计，可得到相对累计频数（relative cumulative frequency）。将互相排斥情形的频数无遗漏地列在一起，便可获得频数表（frequency table）。当样本含量较大（如 100 例以上）时，由频数表与直方图（histogram）可以直观地反映区间数据的分布情况。

（1）编制频数表的方法

例 7-1：某地随机抽取正常成年男子 130 名，其红细胞计数值（10^{12}/L）如图 7-5（也见光盘文件“07-01 定量数据的统计描述.xlsx”，）所示。试对该数据编制频数表并绘制直方图。

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	在A～Z列范围内输入或复制粘贴数据，然后在“频数表”与“直方图”查看结果												
2	4.54	4.31	5.31	4.52	5.67	3.87	3.22	5.11	4.75	5.99	4.97	4.41	4.40
3	3.97	3.73	4.79	4.50	3.77	4.22	5.55	4.90	5.19	5.55	5.07	5.20	5.20
4	4.85	3.64	5.37	5.49	4.88	3.83	3.97	4.16	5.21	5.47	4.83	4.79	5.18
5	5.45	4.14	4.62	4.66	5.23	4.50	4.33	4.57	4.34	4.78	4.11	4.36	4.97
6	5.41	4.26	4.26	4.60	5.82	4.70	5.15	4.79	4.17	4.71	5.43	5.47	5.11
7	5.72	3.48	5.33	4.41	4.66	4.73	4.98	5.03	5.35	4.97	4.53	3.69	5.66
8	3.44	4.38	4.37	5.85	4.41	4.52	5.22	4.79	4.01	4.70	4.22	4.99	4.89
9	4.57	4.48	4.98	5.21	5.10	5.98	5.06	4.18	3.81	4.10	4.23	4.64	5.05
10	5.35	4.79	5.03	6.09	4.49	3.70	3.91	5.80	5.12	3.68	4.46	4.73	5.78
11	4.08	4.5	4.53	4.33	5.15	4.28	4.06	4.99	5.08	5.19	4.45	4.33	4.52

图 7-5　红细胞计数资料

解：编制频数表的步骤如下：

①求全距（range）（也叫极差）（见工作表“频数表”单元格 C5）。全距=最大值−最小值，即 R=6.09−3.22=2.87。

②确定组距（class width）。通常分 8～15 个组，为方便计算，取全距的十分之一，再略加调整。组距= R /10=2.87/10=0.287≈0.30，实取组距为 0.30。

③列出组段（class intervals）。首先取比最小值 3.22 略小一点的数 3.20 为第 1 组段的起点（下限），记为“3.20～”；第 2 组的下限为 3.20+0.30 = 3.50，记为“3.50～”；…；最后一个组段必须包含最大值 6.09，本例记为“5.90～6.20”。

Excel 产生频数表与直方图时要求有上限值。为了程序的通用性，这里采用每一组别的下限值与组距之和，减去一个十分小的数（这里设定为 10^{-10}），求该组段的上限值。第 1 组段的上限值取 3.5（=3.20+0.3+10^{-10}），第 2 组段的上限值取 3.8（=3.50+0.3+10^{-10}），…，其他组段的上限值依此类推（见表 7-3）。

④划记归组获得各组段频数。因为“正”为五画，手工制作频数表时，通常以此划记计数。

⑤求相对频数、累计频数、相对累计频数。

为直观反映频数表的结果，可利用直角坐标系绘制频数或相对频数的直方图（histogram），横轴表示变量的“各种情形”，即组段；纵轴表示频数或相对频数（频率）。本例的频数表和直方图分别见表 7-3 和图 7-6。

结果分析：由表 7-3（频数表）和图 7-6（直方图）可见，130 名正常成年男子的红细胞计数的分布呈中间高、两边低、左右基本对称，大多数人的红细胞计数值在 4.40~5.00（10^{12}/L），两个极端组 3.20～与 5.90～各占 3 人。表明数据近似呈正态分布。

表 7-3　130 名正常成年男子红细胞计数（10^{12}/L）

组段 （1）	上限值 （2）	频数 （3）	累计频数 （4）	相对频数(%) （5）	相对累计频数（%） （6）
3.2～	3.50	3	3	2.31	2.31
3.5～	3.80	6	9	4.62	6.92
3.8～	4.10	9	18	6.92	13.85
4.1～	4.40	20	38	15.38	29.23
4.4～	4.70	24	62	18.46	47.69
4.7～	5.00	24	86	18.46	66.15
5.0～	5.30	21	107	16.15	82.31
5.3～	5.60	13	120	10.00	92.31
5.6～	5.90	7	127	5.38	97.69
5.9～	6.20	3	130	2.31	100.00
合计		130		100.00	

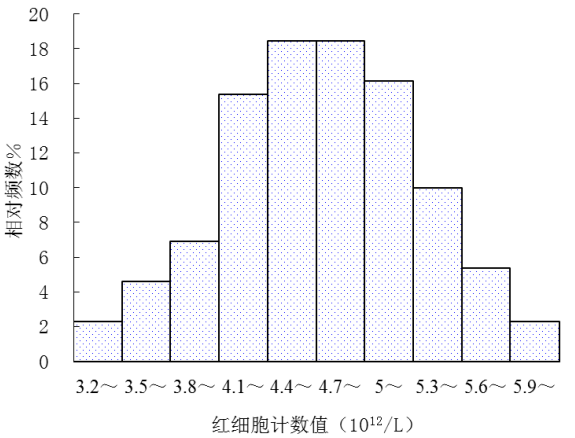


图 7-6　130 名正常成年男子的红细胞计数（10^{12}/L）

（2）FREQUENCY 函数编制频数表

采用 Excel 的 FREQUENCY 函数编制频数表，可按下列步骤进行。

①将数据存放在“数据”工作表的单元格区域 A2:Z3000 范围以内。

②计算最大值与最小值，按上述频数表制作方法划分组段，确定每个组段的“上限值”。确定上限值是 Excel 编制频数表的关键，将这些上限值存放在“频数表”工作表的单元格区域 G2:G30（Excel 称这组上限值所在区域为接收区域（bins array））。

③选取结果存放的单元格区域 H2:H30。

④在编辑栏输入公式“=FREQUENCY(数据!A2:Z3000,频数表!G2:G30)”，按

Ctrl+Shift+Enter 组合键（不要单击“确定”按钮），即可获得各组段相应的频数，结果见表 7-3 的第（3）栏。



注意：因为所求各组段相应的频数是一组数据，占据了多个单元格，为了一次同时获得多个结果，必须采用 Ctrl+Shift+Enter 组合键，否则计算错误。

为了获得表 7-3 的第（4）栏累计频数，先让单元格 J2 等于单元格 H2。在单元格 J3 处输入公式“=J2+H3”，将此公式向下拖放至单元格 J30。

为了获得第（5）栏的相对频数，先在 C13 单元格键入公式“=COUNT(数据!A1:Z3000)”，求“数据”工作表的单元格区域 A2:Z3000 范围以内的数据个数。在 K2 单元格键入公式“=H2/$C$13*100”，单击“确定”按钮，将鼠标移至 K2 单元格的填充柄处，向下拖放该公式至 K30。

为了获得第（6）栏的相对累计频数，在 L2 单元格键入公式“=J2/$C$13*100”，单击“确定”按钮，将鼠标移至 L2 单元格的填充柄处，向下拖放该公式至 L30。

温馨提示：已编制好的程序在所配光盘中，文件名为“07-01 定量数据的统计描述.xlsx”。使用方法是：将需要制作频数表和直方图的数据复制、粘贴或直接输入到“数据”工作表的单元格区域 A2:Z3000 范围以内（注：这一范围是作者编制程序时人为规定的）。在“频数表”工作表中，根据单元格 C7 中的数值（全距/10）确定组距（即修改单元格 C9）；根据单元格 C2 中的数值（最小值）确定第 1 组段下限值（即修改单元格 E2）。

相同的数据，不同的组距（修改单元格 C9）或不同的第 1 组段下限值（修改单元格 E2），将得到不同的频数表。

（3）直方图

直方图可以更直观地反映数据的分布情况。

“07-01 定量数据的统计描述.xlsx”文件中，“直方图”工作表会自动根据“频数表”结果绘制直方图。这个工作表自动生成两个直方图，上面一个图是未经修改的图形，下面一个图是按下列方式修改后的图形。

温馨提示：

①根据你的数据，修改横坐标标题，如本例的横坐标标题修改为“红细胞计数值（10^{12}/L）”。

②根据 A、B、C 三列数据的实际范围修改直方图的数据范围。即在图上单击鼠标右键，在弹出的对话框选取“选择数据”，根据实际数据范围修改“图表数据区域”中“=直方图!$A$4:$A$13,直方图!$C$4:$C$13”的两个“13”。

③如果想绘制“频数”的直方图，可将 C 列数据更改为 B 列数据，即将“图表数据区域”中“=直方图!$A$4:$A$13,直方图!$C$4:$C$13”中的“!$C$4:$C$13”修改为“!$B$4:$B$13”。

（4）数据分析工具：直方图

Excel 数据分析工具中的“直方图”不仅可获得 FREQUENCY 函数得到的各组段频数，还可获得直方图。为了建立例 7-1 的频数表，并绘制直方图，可按如下步骤进行操作。

①将数据存放在单元格区域 A2:M11。参见光盘文件“07-02 工具直方图.xlsx”。

②计算最大值与最小值，按上述“频数表制作方法”划分组段，确定每个组段的“上限值”。将这些上限值存放在单元格区域 B14:B23（Excel 称为接收区域）。

③单击“数据→分析→数据分析”，可弹出“数据分析”对话框。从中选择直方图，弹出“直方图”对话框。

④在“输入区域”用鼠标选取单元格区域 A2:M11；在“接收区域”用鼠标选取“B14:B23”；在输出区域中选取单元格 C13，并复选“图表输出”（见图 7-7）。



图 7-7 “直方图”对话框

单击“确定”按钮后，得到如图 7-8 所示的结果。其中“接收”和“频率”分别相当于表 7-3 的第（2）栏“上限值”和第（3）栏“频数”。图 7-8 的直方图图形只是一个“草图”，为了修改成图 7-6 的格式，要做如下工作：

①将表 7-3 的第（1）栏“组段”的内容复制、粘贴到单元格区域 C14:C23。因为“其他”的频率为 0，所以将第 24 行整行删除。

②在图例上单击鼠标右键，删除图例“频率”；在图表标题上单击鼠标右键，删除图表标题“直方图”；设置横坐标轴标题为“红细胞计数值（10^{12}/L）”，设置纵坐标轴标题为“频数”。

③将鼠标移至任意一个直条上，单击鼠标右键选择“设置数据系列格式”，选择右上角处的“系列选项”标签，调整分类间距为 0%，见图 7-9。

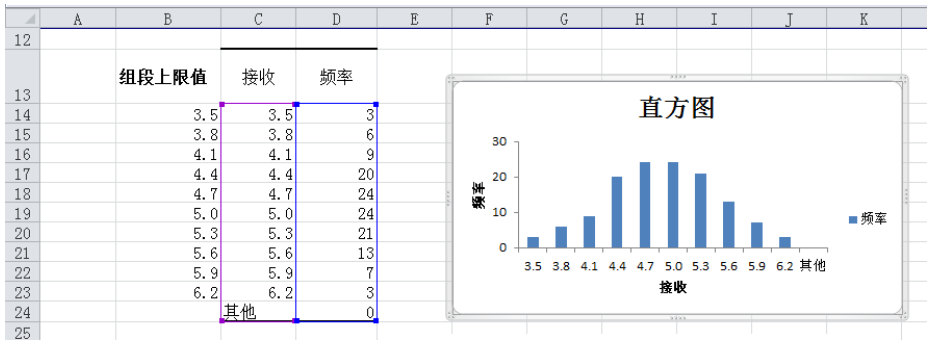


图 7-8　由分析工具“直方图”获得的结果

④继续在“设置数据系列格式”对话框左侧选择“填充”，选取“填充效果”，调整前景色、背景色、图案，单击“关闭”按钮。

⑤调整图形的大小和字号。如果需要还可分别在横轴、纵轴处单击鼠标右键，调整横轴、纵轴的坐标轴格式等。

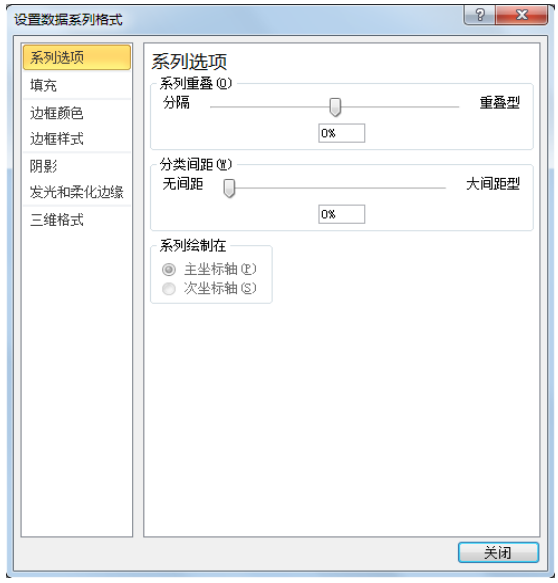


图 7-9　直方图直条间距宽度的调整

（5）直方图的形状

从直方图可以获取区间数据的分布信息，如直方图可以反映数据的分布呈对称（symmetry）还是呈偏态（skewness）。

如果在直方图的中心画一条垂线，垂线两边的图形形状和大小相同，就说直方图对称，

其分布为对称分布，参见图 7-6、图 7-10a、图 7-10b。其中图 7-6 近似服从正态分布（normal distribution）。

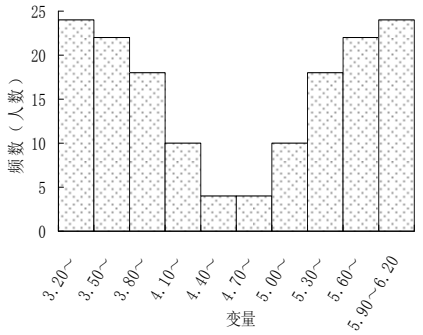


图 7-10a　对称直方图

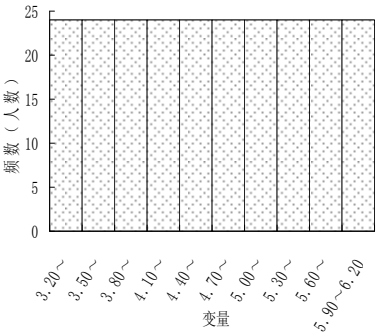


图 7-10b　对称直方图

根据直方图中拖出的长尾所在的方向，偏态分布可分为左偏态和右偏态。在直角坐标系中，原点 0 以左为负，原点 0 以右为正，所以左偏态也称为负偏态（见图 7-11a），右偏态也称为正偏态（见图 7-11b）。

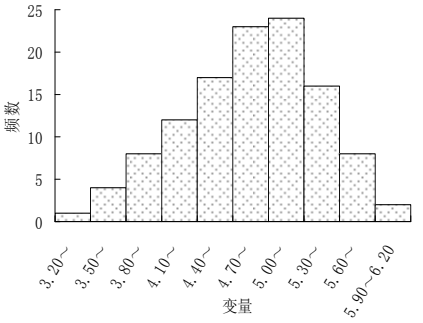


图 7-11a　负（左）偏态直方图

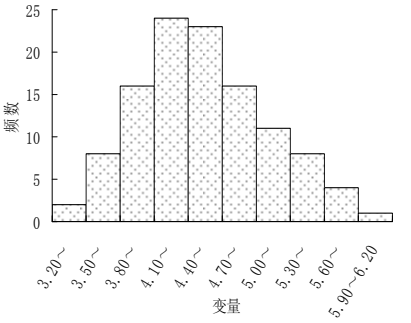


图 7-11b　正（右）偏态直方图

7.3.2　集中位置指标

一般从集中趋势（central tendency）与离散趋势（tendency of dispersion）两个方面分别描述区间数据的集中位置（即平均水平）与离散程度（即变异程度）。尽管直方图可以直观反映区间数据的这两种趋势，但为了用确切数值表示，还需要分别采用集中位置指标和离散程度指标。

例 7-2：测得 18 名健康成年男子的脉搏（次/分）值分别为：75、72、66、79、65、

70、75、67、75、64、70、73、73、72、71、79、69、60，将这些数据值存放在 Excel 的第 A 列（见光盘文件“07-03 集中位置与离散程度.xlsx），试计算该数据的集中位置指标和离散程度指标。

反映集中位置（central position）的指标统称为平均数（average）。平均数包括：算术均数、中位数、几何均数、众数、调和均数、修剪均数等。在日常工作中，前 3 种平均数更常用，特别是算术均数。

（1）算术均数

算术均数（arithmetic mean），简称均数（mean）。总体均数用希腊字母μ（读作 mu）表示，样本均数用 $\bar{X}$ 表示。

总体均数与样本均数的计算公式分别为：

$$\mu=\left(\sum_{i=1}^{N}X_i\Big/N\right)=(x_1+x_2+\cdots+x_N)/N$$

$$\bar{X}=\left(\sum_{i=1}^{n}X_i\Big/n\right)=(x_1+x_2+\cdots+x_n)/n$$

式中，Σ（大写希腊字母，读作 sigma）为求和符号，X_i 为第 i 个观察值，N 为总体总例数，n 为样本总例数。对于例 7-2，样本均数为：

$$\bar{X}=\frac{75+72+\cdots+60}{18}=70.83\ \text{（次/分）}$$

适用条件：随机变量 x_i 的数据满足对称分布或正态分布。

Excel 计算公式：“=AVERAGE(A:A)”。

均数的分子为“求和（sum）”，分母为“计数（count）”，可分别采用 Excel 公式：“=SUM(A:A)”、“=COUNT(A:A)”进行计算。



注意：

①例 7-2 数据实际存放在 Excel 的单元格区域 A2:A19，但为了计算公式的通用性，单元格区域采用了 A:A 替代，即对整个第 A 列进行计算。

②均数受极端值的影响较大。

（2）中位数

将一组数据从小到大排列，位置居中的那个数值称为中位数（median，M）。

当样本含量 n 为奇数时，位置居中的那个数值 $x_{(n+1)/2}$ 就是中位数；当 n 为偶数时，位置居中的两个数值的平均 $(x_{n/2}+x_{(n/2)+1})/2$ 就是中位数。例 7-2 的样本含量为偶数，位置居中的两个值分别为第 18/2＝9 位的 71 和第 10 位的 72，因此中位数等于 (71+72)/2=71.5（次/分）。

适用条件：随机变量 x_i 的数据呈明显的偏态分布，或者数据的分布不明确，或者某些数据没有具体的取值（如“≤20”、“>50”）。

Excel 计算公式：“=MEDIAN(A:A)”。



注意：

对于正态分布或对称分布的数据，理论上中位数等于均数。中位数的缺点是：没有充分利用观察到的每个变量值的信息，只利用了“位置居中”的其中一个或两个值。

（3）几何均数

将 n 个观察值 X_i 相乘的乘积再开 n 次方，所得的根就是几何均数（geometric mean，G）。若对各观察值 X_i 取对数，对数值 $\log X_i$ 的算术均数的反对数即为 G。计算公式：

$$G=\sqrt[n]{x_1x_2\cdots x_n}=\log^{-1}\left(\frac{\log x_1+\log x_2+\cdots+\log x_n}{n}\right)=\log^{-1}\left(\frac{\sum_{i=1}^{n}\log X_i}{n}\right)$$

对于例 7-2（假定服从对数正态分布），有：

$$G=\sqrt[18]{75+72+\cdots+60}=70.66\text{（次/分）}$$

适用条件：随机变量 x_i 的数据服从对数正态分布，如抗体的滴度、药物的效价等。

Excel 计算公式：“=GEOMEAN (A:A)”。



注意：

①这里的对数（反对数）是以 10 为底的常用对数，如果采用以 e=2.71828 为底的自然对数，或采用以其他数字（如 2）为底的对数，将获得完全相同的几何均数。

②同一组数据若服从正态分布，理论上均数与中位数相等，但几何均数小于均数。

（4）众数

众数（mode）是一组观察值中出现频数（或次数）最多的值，直方图中直条最高的那个组段即为众数所在组段。

对于例 7-2，出现频率最高的观察值为 75（次/分）（共出现 3 次）。

Excel 计算公式：“=MODE (A2:A60000)”。



注意：

①如果数据呈正态分布，则有众数=均数=中位数。

②众数的概念简单，但若出现若干个最高相同频数时，则众数有多个观察值。

③众数也没有充分利用样本观察值的全部信息。

④例 7-2 数据实际存放在 Excel 的单元格区域 A2:A19，但为了计算公式的通用性，单元格区域采用了 A2:A60000 替代（因为 A1 为文字标题，所以计算众数时采用单元格区域 A:A 不适合，此时屏幕结果单元格将显示“#NUM!”）。

（5）调和均数

调和均数（harmonic average）是倒数的算术平均数的倒数。

计算公式：

$$H=\frac{1}{\frac{1}{n}(\frac{1}{x_1}+\frac{1}{x_2}+\cdots+\frac{1}{x_n})}=\frac{n}{\sum_{i=1}^{n}\frac{1}{X_i}}$$

例 7-2 的脉搏（次/分）$H=\dfrac{1}{\frac{1}{18}(\frac{1}{75}+\frac{1}{72}+\cdots+\frac{1}{60})}=70.48$。

Excel 计算公式：“= HARMEAN (A:A)”。

（6）修剪均数

通常情况下，最大的几个值或最小的几个值不稳定，为了获得更加稳健的平均值，可按一定百分数（或比例）修剪（经排序后的）数据首尾两端的观察值，然后再求平均。这样获得的均数就是修剪均数（trim average）。如果按 2/18 修剪，即从例 7-2 数据中去掉最大值与最小值（79 与 60），再求平均得到 71（次/分）。

Excel 计算公式：“=TRIMMEAN(A2:R2,2/18)”。

以上的几个平均数中，以均数最常用；但数据如果呈偏态，或者两端值波动较大时，也常采用中位数。

7.3.3　离散程度指标

由于个体差异的存在，获得的个体观察值之间有所不同。反映观察值之间差异大小的指标就是离散程度（dispersion）指标，或称之为变异度（variability）指标。

设有甲、乙、丙三名医生，分别对 5 份血样进行红细胞计数（万/mm^3），甲得出了 560、540、500、460、440，乙得出了 520、510、500、490、480，丙得出了 510、505、500、495、490，如图 7-12 所示。

结果分析：甲、乙、丙三名医生所得 5 个血样结果的均数均为 500，之和均为 2500，但甲医生得出的 5 个观察值间的差异（离散程度）较大，而丙医生得出的 5 个观察值间的差异较小。由此可见，为了全面描述一组区间数据的统计学特征，不仅需要计算集中位置

指标，而且还需要计算离散程度指标。

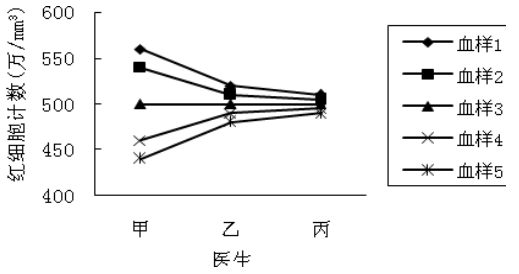


图 7-12　三名医生对 5 份血样的红细胞计数结果

反映离散程度的常用指标有：全距、四分位数间距、方差、标准差和变异系数。

1. 全距

全距（range，R），也叫极差，是最大值（maximum）与最小值（minimum）之差。

全距容易理解、计算简单，但其缺点是：

① 一组观察数据从小到大排列后，全距仅利用了两个端点值，未利用全部观察值的信息。

② 观察的样本含量越大，数值波动的范围越大，全距就越大，故样本含量相差悬殊时，不能通过比较全距来比较离散趋势。

③ 若一组观察值的一端或两端有不确切的数值时，不能计算全距。

对于例 7-2，最大值为 79（次/分），最小值为 60（次/分），全距为 79-60=19（次/分）。

Excel 计算公式：“=MAX(A:A)–MIN(A:A)”。

全距越大，数据的离散程度越大。

2. 四分位数间距与箱图

（1）百分位数

百分位数（percentile）是把一组数据从小到大排列，分割成 100 等份，每等份含 1%的观察值，分割界点值就是百分位数，用符号 P_x 表示。P_x 将数据分成两部分，有 x%的数值小于 P_x，(100–x)%的数值大于 P_x。第 50 百分位数为中位数，用 P_{50} 表示。第 0 百分位数为最小值，第 100 百分位数为最大值。第 5、25、75、95 百分位数分别记为 P_5 、P_{25}、P_{75}、P_{95}。

如果有 n 例，则第 x 百分位数是位于第

$$L_x = (n+1) \times x\%$$

位置的观察值。如例 7-2 的 n=18，按小到大顺序排列后的 18 个数值为：

秩次	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
数据	60	64	65	66	67	69	70	70	71	72	72	73	73	75	75	75	79	79

第 25 百分位数（P_{25}）应是位于第

$$L_{25}=(18+1)\times\frac{25}{100}=4.75$$

位置的观察值。即为秩次在 4~5 之间的观察值，即第 25 百分位数应是 66~67 之间的数。

采用内插法，取第 L_x 整数位置值，然后加上 L_x 小数值与[第(L_x＋1)整数位置值－第 L_x 整数位置值]的乘积，即为第 x 百分位数。

第 25 百分位数=66+0.75×(67-66)=66.75。

Excel 计算公式：“= PERCENTILE (A:A,0.25)”。（Excel 结果为 67.5，可能其算法与此有所不同。）

对于 P_{25}、P_{50}、P_{75}、P_{95}，应将 Excel 公式中的 0.05 分别修改为 0.25、0.5、0.75、0.95。由例 7-2 计算 P_5、P_{50}、P_{75}、P_{95} 分别获得 60、71.5、75、79（次/分）。（Excel 结果为 63.4、71.5、74.5、79。）

对于 P_0、P_{25}、P_{50}、P_{75}、P_{100} 五个百分位数可采用 Excel 的四分位数函数 QUARTILE 计算，如 P_0 的 **Excel 计算公式**：“= QUARTILE (A:A,0)”；计算 P_{25}、P_{50}、P_{75}、P_{100} 时，只需将前面 Excel 公式中的“0”参数分别修改为“1”、“2”、“3”、“4”即可。P_{25}、P_{50}、P_{75} 分别成为第 1 四分位数（记为 Q_1）、第 2 四分位数（记为 Q_2）、第 3 四分位数（记为 Q_3）。

Excel 也提供了计算第 k 较小值（从小到大顺序排列后，排在第 k 位置的观察值）和第 k 较大值（从小到大顺序排列后，排在倒数第 k 位置的观察值）的计算函数，对于本例，第 3 较小值的 **Excel 计算公式**：“= SMALL (A:A,3)”，结果为 65；第 3 较大值的 **Excel 计算公式**：“= LARGE (A:A,3)”，结果为 75。

温馨提示：这些计算结果与 SAS、SPSS 软件略有所不同。

（2）四分位数间距

四分位数间距（interquartile range，Q），反映了位于中间的 50%观察值的离散程度。是第 3 四分位数（Q_3）与第 1 四分位数（Q_1）间的差值，即：

$$Q=Q_3-Q_1=P_{75}-P_{25}$$

该指标比全距稍稳定，但仍未考虑每个观察值。

本例的 Q=74.5−67.5=7（次/分）。

Excel 计算公式：“= PERCENTILE (A:A,0.75)– PERCENTILE (A:A,0.25)”。

四分位数间距越大，表示第 3 四分位数与第 1 四分位数间相距较远，离散程度较大。

（3）箱图

可采用箱图（boxplot）更加直观地反映四分位数的分布情况，本例的箱图见图 7-13（绘制方法参见第 6 章）。箱图的两端分别表示最小值（左）与最大值（右），中间箱的左端表

示第 1 四分位数（Q_1），右端表示第 3 四分位数（Q_3），箱的中部直线表示中位数（Q_2，即 P_{50}），加号表示均数。

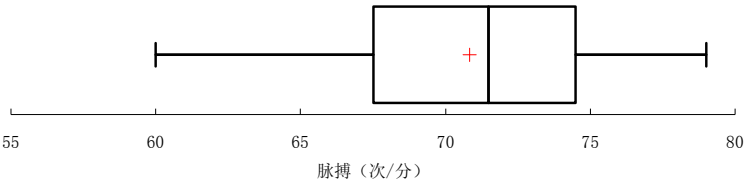


图 7-13　箱图

3. 方差

若要克服全距和四分位数间距不能利用每个观察值的缺点，可采用每个观察值 x_i 与均数 $\overline{x}$ 间差值的绝对值的平均水平，即 $\left(\sum_{i=1}^{n}|X_i-\overline{X}|\right)\Big/n$，来反映离散程度的大小，这称为平均绝对离差。平均绝对离差的 **Excel 计算公式**：“=AVEDEV(A:A)”，本例计算结果为 3.96（次/分）。该指标应用较少。

使用更多的离散指标是方差（variance）。总体方差用 σ^2 表示（σ为小写希腊字母，读作 sigma），计算公式为：

$$\sigma^2=\frac{\sum_{i=1}^{N}(X_i-\mu)^2}{N}$$

由样本获得的方差就是样本方差，用 S^2 表示，公式为：

$$S^2=\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{n-1}$$

该公式的分子部分称为离均差平方和（sum of square），本例的 **Excel 计算公式**：“=DEVSQ(A:A)”，获得结果 438.5（次/分）^{2}；分母部分为自由度(degree of freedom)。

本例的样本方差等于 25.79（次/分）^{2}，**Excel 计算公式**：“= VAR (A:A)”。

方差越大，一组数据的离散程度越大；方差越小，一组数据的离散程度越小。但由于方差的意义相当于离均差平方和的平均值（也称之为均方，mean square），所以度量衡单位也是原数据单位的平方，故难以解释。

4. 标准差

标准差（standard deviation）是方差的平方根。度量衡单位与原变量及其均数的单位相同。该指标是描述正态分布数据最常用的离散程度指标。标准差越大，说明观察值的离散程度越大；标准差越小，说明观察值的离散程度越小。

Excel 计算公式：“=STDEV(A:A)”，本例获得的样本标准差为 5.08（次/分）。

标准差的意义和用途如下：

①说明一组数据的离散程度，若比较样本的度量衡单位相同，均数接近，则标准差较小者观察值离散程度较小，标准差较大者观察值离散程度较大。

②若比较样本（通常有两组或两组以上）的度量衡单位不同，均数相差悬殊，则需利用标准差与均数一起计算变异系数（coefficient of variation，*CV*），计算公式为：

$$CV = \frac{S}{\bar{X}} \times 100\%$$

Excel 计算公式：“=STDEV(A:A)/AVERAGE(A:A)”。本例计算得 *CV*＝0.07。*CV* 越大者离散程度越大。

③用标准差与样本含量的平方根之比可计算标准误（standarded error），计算公式为：

$$S_{\bar{X}} = \frac{S}{\sqrt{n}}$$

Excel 计算公式：“=STDEV(A:A)/COUNT(A:A)^0.5”。由本例计算得到标准误 $S_{\bar{X}} = \frac{5.08}{\sqrt{18}} = 1.20$。

标准误反映了样本均数的离散程度，衡量了抽样误差的大小，在第 10 章的置信区间估计与假设检验中，将要详细说明。标准误越小，表示样本均数与总体均数越接近，即用样本均数估计总体均数的可靠性越大；抽样误差大，样本均数离总体均数越远，用样本均数估计总体均数的可靠性越小。

④ 结合均数与直方图解释标准差。

如果一组数据近似服从正态分布，则在均数的 1 个标准差范围（即 $\bar{X} \pm S$）内，大约包括 68%的个体观察值；在均数的 2 个标准差范围（即 $\bar{X} \pm 2S$）内，大约包括 95%的个体观察值；在均数的 3 个标准差范围（即 $\bar{X} \pm 3S$）内，大约包括 99.7%的个体观察值。如图 7-14 所示。

利用均数与标准差可计算参考值范围的大小。如利用 $\bar{X} \pm 1.96S$ 可计算 95%参考值范围。本例该组数据（假定满足正态分布条件）的 95%参考值范围为（60.88, 80.79）。

应特别注意，此处的样本含量 *n*=18 较小（比如小于 50），因此本例的结果不可靠。按正态分布计算参考值范围，一般样本含量最好在 100 例以上。

5. 偏度与峰态

（1）偏度

偏度（skewness）反映一组数据的频数分布曲线的高峰是偏左、居中还是偏右。偏度 g_1 的计算公式为：

$$g_1=\frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{S}\right)^3$$

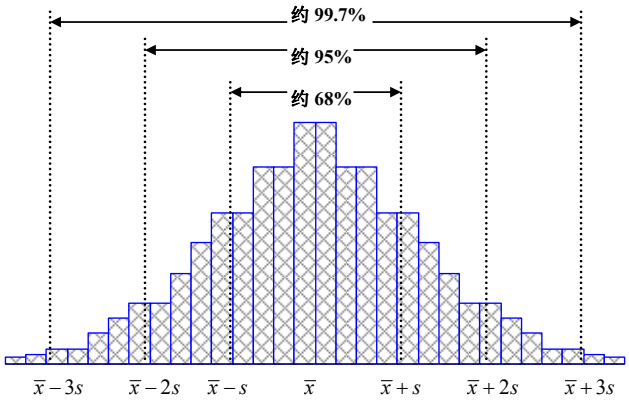


图 7-14　结合均数与直方图解释标准差

式中，X_i 为第 i 个体的观测值，n、$\bar{X}$ 、S 分别为样本含量、样本均数和样本标准差。$g_1<0$ 为负偏峰，表示曲线高峰偏右，长尾向左（负）方向延伸；$g_1=0$ 表示曲线对称；$g_1>0$ 为正偏峰，表示曲线高峰偏左，长尾向右（正）方向延伸。本例有：

$$g_1=\frac{18}{(18-1)(18-2)}\left[\left(\frac{75-70.83}{5.08}\right)^3+\left(\frac{72-70.83}{5.08}\right)^3+\cdots+\left(\frac{60-70.83}{5.08}\right)^3\right]=-0.32$$

获得的值是负数，为负偏峰，曲线高峰偏右，长尾向左（负）方向延伸（见图 7-13 的箱图）。

Excel 计算公式：“=SKEW(A:A)”。

（2）峰态

峰态（kurtosis）反映一组数据的频数分布的高峰是平阔峰、正态峰还是尖峭峰。计算公式为：

$$g_2=\frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{S}\right)^4-\frac{3(n-1)^2}{(n-2)(n-3)}$$

$g_2<0$，表示曲线高峰平坦，为平阔峰；$g_2=0$，表示曲线峰度适中，为正态峰；$g_2>0$，表示曲线高峰尖且陡峭，为尖峭峰。本例有：

$$g_2=\frac{18(18+1)}{(18-1)(18-2)(18-3)}\left[\left(\frac{75-70.83}{5.08}\right)^4+\left(\frac{72-70.83}{5.08}\right)^4+\cdots+\left(\frac{60-70.83}{5.08}\right)^4\right]-$$

$$\frac{3(n-1)^2}{(n-2)(n-3)}=-0.12$$

本例获得的值是一个很小的负数，为平阔峰，曲线高峰平坦。

Excel 计算公式：“=KURT(A:A)”。

程序修改说明：

在光盘中，文件名为：“07-03 集中位置与离散程度.xlsx”的 Excel 程序可完成集中位置指标（7.3.2 节）和离散程度指标（7.3.3 节）的计算，并可自动产生类似于图 7-13 的箱图。

操作方法是：将一组区间数据输入或复制、粘贴在第 A 列单元格区域（从第 A2 单元格开始）中。所有集中位置指标和离散程度指标的计算结果便随即产生，并自动产生箱图。

为了计算满足各自的需要，还可作如下几点修改：

① 为了计算修剪均数，可在单元格 D16 中给出修剪百分数（比例）（0~1 间的数值）。

② 为了计算第 x%分位数，可在单元格 J11 中给出 x%（0~1 间的数值，如 95%=0.95）。

③为了让“箱图”工作表中的箱图更加好看，可对第二个箱图进行修改，如将横坐标标题“变量”修改为“脉搏（次/分）”，横坐标的刻度 0~90 修改为 55~80。

光盘文件“07-01 定量数据的统计描述.xlsx”同样具有“07-03 集中位置与离散程度.xlsx”文件的计算功能，并且适用于多列数据的计算（后一文件只适用于一列数据的计算）。

7.3.4　数据分析工具：描述统计

由“描述统计（descriptive statistics）”数据分析工具可一次性产生以上介绍的大多数指标。具体操作步骤为：

将所要计算的一组区间数据存放在第 A 列（A1 单元格为变量名），单击“数据→分析→数据分析”，可弹出数据分析对话框。从中选择“描述统计”，将出现“描述统计”对话框（见图 7-15）。



图 7-15　“描述统计”对话框

在输入区域用鼠标选取第 A 列“$A:$A”。因为 A1 单元格为变量名称，所以在“标志

位于第一行”前打钩（否则会提示出错）。将该工具输出产生的结果放在工作表的 B2 单元格及其右下侧，输出结果也可存放在新的工作表或新工作簿中。

“汇总统计”是必选项，只有选择了此选项，方可获得均数、中位数、众数 3 个集中趋势指标，以及全距、样本方差、标准差、偏度、峰度 5 个离散趋势指标。此外，也给出了最小值、最大值、总和、样本观测例数 4 个基本数据，以及标准误。

选择“平均数置信度”，在其后给出置信度，如 95%，可获得 t 界值与标准误的乘积值，即均数估计值的误差范围（bound on the error of estimation），也就是公式 $\bar{X} \pm t_{\alpha/2,\nu} S_{\bar{X}}$ 的后半部分，即 $t_{\alpha/2,\nu} S_{\bar{X}}$，如需获得置信区间（confidence interval），则需采用均数与该值相减或相加。相减获得置信区间下限值，相加获得置信区间上限值。

如需获得第 K 个较大值或较小值，可在“描述统计”对话框中选取。

填完如图 7-15 所示的“描述统计”对话框后，单击“确定”，可获得如图 7-16（见光盘文件“07-04 工具描述统计.xlsx）所示的中间部分（第 B、C 列）的结果。因为 Excel 输出的统计学名词，有几个与一般生物统计学教材有些不同（尽管 Office 2003 已做了部分修改），所以在图 7-16 右侧给出了每个 Excel 统计学名词的中英文对照表，第 G 列是采用 Excel 函数计算得到的结果。读者可从该光盘文件理解每个指标的 Excel 函数计算方法。

	A	B	C	D	E	F	G
1	脉搏（次/分）						
2	75	脉搏（次/分）					
3	72					说 明	
4	66	平均	70.83333333		均数	Mean	70.83333333
5	79	标准误差	1.19708251		标准误	Standard Error	1.19708251
6	65	中位数	71.5		中位数	Median	70.65810302
7	70	众数	75		众数	Mode	75
8	75	标准差	5.078790963		标准差	Standard Deviation	5.078790963
9	67	方差	25.79411765		方差	Sample Variance	25.79411765
10	75	峰度	-0.11837		峰度	Kurtosis	-0.11837
11	64	偏度	-0.321108665		偏度	Skewness	-0.321108665
12	70	区域	19		全距	Range	19
13	73	最小值	60		最小值	Minimum	60
14	73	最大值	79		最大值	Maximum	79
15	72	求和	1275		总和	Sum	1275
16	71	观测数	18		样本含量	Count	18
17	79	置信度(95.0%)	2.525626855		误差范围	Bound on the error of estimation	2.525623328
18	69						
19	60						

图 7-16　描述统计的计算结果及有关说明

7.4　名义数据的统计描述

在 5.2.4 节，采用“数据透视表”，由原始数据获得了列联表数据，这种列联表数据也就是计数资料。此外，也可采用 COUNTIF 函数获得一维列联表。

7.4.1　COUNTIF 函数清点计数

COUNTIF 函数的格式为：COUNTIF(数据范围,计数条件)。比如清点例 5-3 数据（即“03-04 妊娠分娩.xlsx 文件）中 1402 名分娩产妇“文化程度”为“大学”者有多少，可采用 Excel 计算函数为：

“=COUNTIF(G2:G1403,"大学")”

所得结果为：1402 名分娩产妇中“文化程度”为“大学”者有 208 人；“G2:G1403”为文化程度变量所在的单元格区域，因为“大学”为文本数据，所以应用英文引号引起来。采用 Excel 函数“=COUNTIF(G2:G1403,"文盲")”可计算获得 1402 名分娩产妇中“文化程度”为“文盲”者有 41 人；其他“文化程度”类别的计数可仿照以上函数计算。这些结果与图 5-19 相同。

7.4.2　率的标准化法

采用统一的标准对内部构成不同的各组率进行调整的方法就是率的标准化法。进行率的比较时，应注意各组的内部构成是否相同。若存在差异而又需要比较时，需要利用标准化法对率进行调整。调整后的率称为标准化率（standardized rate）或调整率（adjusted rate）。

例 7-3：A、B 两县男性的年龄组肝癌死亡率资料见表 7-4，试对两县的死亡率作标准化。

表 7-4　A、B两县男性的年龄组肝癌死亡率

年龄组（1）	A 县				B 县			
	死亡数（2）	人口数（3）	人口构成比（4）	死亡率（5）	死亡数（6）	人口数（7）	人口构成比（8）	死亡率（9）
0～	**26**	**363726**	0.6555	0.000071	**24**	**409698**	0.6949	0.000059
30～	**77**	**63843**	0.1150	0.001206	**77**	**72273**	0.1226	0.001065
40～	**105**	**47657**	0.0859	0.002203	**106**	**45072**	0.0765	0.002352
50～	**89**	**34282**	0.0618	0.002596	**86**	**32371**	0.0549	0.002657
60～	**71**	**23941**	0.0431	0.002966	**56**	**18209**	0.0309	0.003075
70～	**35**	**21468**	0.0387	0.001630	**24**	**11914**	0.0202	0.002014
合计	403	554917	1.0000	0.000726	373	589537	1.0000	0.000633

解：表 7-4 中死亡数和人口数为已知数据，根据已知数据按图 7-17（也见光盘文件“07-05 率的标准化.xlsx）所给公式，可简单计算人口构成比和死亡率两个相对数。

	A	B	C	D	E	F	G	H	I	J
1										
2	年龄组(1)	A县					B县			
3		死亡数(2)	人口数(3)	人口构成比(4)	死亡率(5)		死亡数(6)	人口数(7)	人口构成比(8)	死亡率(9)
4	0~	26	363726	=C4/C$10	=B4/C4		24	409698	=H4/H$10	=G4/H4
5	30~	77	63843	=C5/C$10	=B5/C5		77	72273	=H5/H$10	=G5/H5
6	40~	105	47657	=C6/C$10	=B6/C6		106	45072	=H6/H$10	=G6/H6
7	50~	89	34282	=C7/C$10	=B7/C7		86	32371	=H7/H$10	=G7/H7
8	60~	71	23941	=C8/C$10	=B8/C8		56	18209	=H8/H$10	=G8/H8
9	70~	35	21468	=C9/C$10	=B9/C9		24	11914	=H9/H$10	=G9/H9
10	合计	=SUM(B4:B9)	=SUM(C4:C9)	=SUM(D4:D9)	=B10/C10		=SUM(G4:G9)	=SUM(H4:H9)	=SUM(I4:I9)	=G10/H10

图 7-17　Excel 计算构成比和死亡率的公式

1. Excel 计算构成比和死亡率

具体操作如下：

①将 A 县死亡数和人口数分别存放在 B4:B9 和 C4:C9，B 县死亡数和人口数分别存放在 G4:G9 和 F4:F9。

②用鼠标单击 B10 单元格，单击“公式→自动求和（图标Σ）”，在编辑栏会自动显示公式“=SUM(B4:B9)”。将该公式向右拖放填充到单元格 C10 和 D10。

③用鼠标单击 D4 单元格，键入公式“=C4/C$10”，向下拖放填充该公式至 D9，获得 A 县人口构成比。

④用鼠标单击 E4 单元格，键入公式“=B4/C4”，向下拖放填充至 E10，获得 A 县死亡率。注意：合计栏的死亡率，也叫平均死亡率，计算需采用合计死亡数除合计人口数。

⑤复制单元格区域 B10:E10，单击 G10 后粘贴；复制单元格区域 D4:E9，单击 I4 后粘贴。由此可简单完成 B 县的有关计算。注意按要求增加或减少小数位数。

2. 率的标准化

从表 7-4 可见，与 B 县相比，A 县人口中的低年龄组（40 岁以下）人口比重较少，而高年龄组（40 岁以上）人口比重较大。A 县男性平均肝癌死亡率（0.000726）高于 B 县（0.000633），而从各年龄组死亡率看不出这种差异。该结果是否由于 A 县年龄偏大造成？因此有必要对该数据的死亡率进行标准化。

率的标准化首先应选择一个有代表性、内部构成相对稳定的“标准人口”，一般可选所有比较组“人口数”之和。由此计算标准人口构成比（如(363726+409698)/(554917+589537) = 0.6758），用“标准人口构成比”乘以“原死亡率”得“分配死亡率”（如 0.6758×0.000071=0.000048），各年龄组分配死亡率之和就是标准化死亡率（如(0.000048+0.000143+…+0.000048) =0.000678），该指标可以相互比较。本例计算得 A、B 两县的标准化死亡率分别为 0.000678、0.000684，即分别为 67.8/10 万、68.4/10 万（见表 7-5）。表明 A 县的男性肝癌死亡率并不高于 B 县，其标准化结果反而较低。

计算获得表 7-5 的 Excel 公式见图 7-18（也见光盘文件“07-05 率的标准化.xlsx”）。单元格区域 C4:C10 与 H4:H10 分别为 A、B 两县男性的人口数。先将计算出的 A、B 两县的

肝癌死亡率复制，分别选择性粘贴（数值）在单元格区域 N4:N10、Q4:Q10；然后分别在 M4、O4、M10 三个单元格建立公式，再采用复制、粘贴其余相应单元格即可。

表 7-5　A、B两县男性的肝癌死亡率的标准化

年龄组（1）	标准人口构成比（2）	A 县		B 县	
		原死亡率（3）	分配死亡率（4）	原死亡率（5）	分配死亡率（6）
0～	0.6758	0.000071	0.000048	0.000059	0.000040
30～	0.1189	0.001206	0.000143	0.001065	0.000127
40～	0.0810	0.002203	0.000179	0.002352	0.000191
50～	0.0582	0.002596	0.000151	0.002657	0.000155
60～	0.0368	0.002966	0.000109	0.003075	0.000113
70～	0.0292	0.001630	0.000048	0.002014	0.000059
合计	1.0000	0.000726	0.000678	0.000633	0.000684

	L	M	N	O	P	Q	R
2	年龄组（1）		A县			B县	
3		标准人口构成比（2）	原死亡率	分配死亡率(4)		原死亡率	分配死亡率(6)
4	0～	=(C4+H4)/($C$10+$H$10)	0.00007148	=$M4*N4		0.00005857	=$M4*Q4
5	30～	=(C5+H5)/($C$10+$H$10)	0.00120608	=$M5*N5		0.00106540	=$M5*Q5
6	40～	=(C6+H6)/($C$10+$H$10)	0.00220324	=$M6*N6		0.00235179	=$M6*Q6
7	50～	=(C7+H7)/($C$10+$H$10)	0.00259611	=$M7*N7		0.00265669	=$M7*Q7
8	60～	=(C8+H8)/($C$10+$H$10)	0.00296562	=$M8*N8		0.00307540	=$M8*Q8
9	70～	=(C9+H9)/($C$10+$H$10)	0.00163033	=$M9*N9		0.00201443	=$M9*Q9
10	合计	=SUM(M4:M9)	0.00072623	=SUM(O4:O9)		0.00063269	=SUM(R4:R9)

图 7-18　率的标准化的 Excel 计算公式

7.4.3　动态数列

动态数列（dynamic series）是一组按照时间顺序（如年份）排列的统计指标（如人均住院费、发病率）。采用这些动态数列，可计算出反映事物或现象的发展水平和发展趋势的有关指标。

例 7-4：某地 1993～2000 年病毒性肝炎的人均住院费见表 7-6，对该数据计算绝对增长量、发展速度、增长速度、平均发展速度、平均增长速度等指标。

表 7-6　某地 1993～2000 年病毒性肝炎人均住院费动态

年份	人均住院费（元）	绝对增长量（%）		发展速度（%）		增长速度（%）		平均发展速度（%）	平均增长速度（%）
		累计	逐年	定基	环比	定基	环比		
1993	1332.50	—	—	—	—	—	—	—	—
1994	3806.18	2474	2474	285.64	285.64	185.64	185.64	285.64	185.64

续表

年份	人均住院费（元）	绝对增长量（%）		发展速度（%）		增长速度（%）		平均发展速度（%）	平均增长速度（%）
		累计	逐年	定基	环比	定基	环比		
1995	9004.63	7672	5198	675.77	236.58	575.77	136.58	259.96	159.96
1996	9697.97	8365	693	727.80	107.70	627.80	7.70	193.79	93.79
1997	9091.50	7759	-606	682.29	93.75	582.29	-6.25	161.62	61.62
1998	8934.19	7602	-157	670.48	98.27	570.48	-1.73	146.31	46.31
1999	9508.48	8176	574	713.58	106.43	613.58	6.43	138.75	38.75
2000	8728.22	7396	-780	655.03	91.79	555.03	-8.21	130.80	30.80

1. 绝对增长量

绝对增长量为两个年份的人均住院费相减所得值。具体分为累计增长量和逐年增长量。确定一个基数，如本例以 1993 年人均住院费为基数，各个年份指标与基数相减可获得累计增长量。如果住院费数据存放在单元格区域 B4:B11，可在单元格 D5 建立 Excel 公式为“=$B5-$B$4”，按住 D5 的右下角的填充柄，向下拖放填充至 D11 即可。

下一年与上一年住院费相减可获得逐年绝对增长量。可在单元格 E5 建立 Excel 公式为“=$B5-$B4”，按住 E5 的右下角的填充柄，向下拖放填充至 E11 即可。

2. 发展速度与增长速度

发展速度与增长速度均为比，具体分为定基比和环比。

其他各年的住院费与 1993 年（基数）相比获得发展速度定基比。可在单元格 G5 建立 Excel 公式为“=$B5/$B$4*100”，按住单元格 G5 的右下角填充柄，向下拖放至单元格 G11 即可。乘以 100 后，使发展水平数据变成百分数。

下一年与上一年住院费相比获得发展速度环比。可在单元格 H5 建立 Excel 公式为“=$B5/$B4*100”，按住单元格 H5 的填充柄，向下拖放至单元格 H11 即可。乘以 100 后，使发展水平数据变成百分数。

增长速度=(发展速度–100)（%）。在单元格 J5 与 K5 分别键入公式“=G5-100”与“=H5-100”，可分别获得增长速度的定基比和环比。

3. 平均发展速度与平均增长速度

平均发展速度可概括某段时期的平均发展水平，是发展速度环比指标的几何均数。令 1993 年住院费（即基数）为 f_0，下一年（即 1994 年）为 f_1，…，f_n。求第 n 年时的平均发展速度，可采用公式 $\sqrt[n]{f_n / f_0}$，如 1993～1995 年间的平均发展速度，有 $n = 1995 - 1993 = 2$，$\sqrt[2]{f_2 / f_0} = \sqrt{9004.63 / 1332.50} = 259.96$。

如果将年份和住院费数据分别存放在单元格区域 A4:A11 和 B4:B11，在单元格 M5 输

入 Excel 公式“=(B5/$B$4)^(1/(A5-$A$4))*100”，可获得第 1 年的平均发展速度。按住单元格 M5 右下角的填充柄，向下拖放填充至单元格 M11，可获得第 *n* 年的平均发展速度。

平均增长速度=(平均发展速度–100)（%）。在单元格 N5 键入公式“=M5-100”，按住单元格 N5 右下角的填充柄，向下拖放填充至单元格 N11，可获得平均增长速度。

该例的动态数列分析的所有 **Excel 计算公式**见图 7-19（也见光盘文件“07-06 动态数列.xlsx”）。

	A	D	E	G	H	J	K	M	N
1									
2	年份	绝对增长量%		发展速度%		增长速度%			
3		累计	逐年	定基	环比	定基	环比	平均发展速度%	平均增长速度%
4	1993	-	-	-	-	-	-	-	-
5	1994	=$B5-$B$4	=$B5-$B4	=$B5/$B$4*100	=$B5/$B4*100	=G5-100	=H5-100	=(B5/$B$4)^(1/(A5-$A$4))*100	=M5-100
6	1995	=$B6-$B$4	=$B6-$B5	=$B6/$B$4*100	=$B6/$B5*100	=G6-100	=H6-100	=(B6/$B$4)^(1/(A6-$A$4))*100	=M6-100
7	1996	=$B7-$B$4	=$B7-$B6	=$B7/$B$4*100	=$B7/$B6*100	=G7-100	=H7-100	=(B7/$B$4)^(1/(A7-$A$4))*100	=M7-100
8	1997	=$B8-$B$4	=$B8-$B7	=$B8/$B$4*100	=$B8/$B7*100	=G8-100	=H8-100	=(B8/$B$4)^(1/(A8-$A$4))*100	=M8-100
9	1998	=$B9-$B$4	=$B9-$B8	=$B9/$B$4*100	=$B9/$B8*100	=G9-100	=H9-100	=(B9/$B$4)^(1/(A9-$A$4))*100	=M9-100
10	1999	=$B10-$B$4	=$B10-$B9	=$B10/$B$4*100	=$B10/$B9*100	=G10-100	=H10-100	=(B10/$B$4)^(1/(A10-$A$4))*100	=M10-100
11	2000	=$B11-$B$4	=$B11-$B10	=$B11/$B$4*100	=$B11/$B10*100	=G11-100	=H11-100	=(B11/$B$4)^(1/(A11-$A$4))*100	=M11-100

图 7-19　动态数列分析的 Excel 计算公式

根据动态数列的发展趋势，可利用数学模型或某种统计方法对未来时期或时点指标做出预测，参见后面的“预测分析”一章。

图 7-17 与图 7-18 所对应的文件中，是按了“Ctrl+`（左上侧 Tab 键之上的键）”组合键后的效果，这个组合键是是否显示公式的开关语句，如果再按“Ctrl+`”组合键，则按常规显示公式所计算的结果。Excel 出现“###”表示显示的单元格宽度不够，增加每列数据的宽度，可消除“###”的显示。

第 8 章

区间数据的假设检验

利用 Excel 的“数据分析”工具，可完成 t 检验、z 检验、F 检验（即方差分析），本章将主要介绍这些假设检验以及相关的统计分析方法。

8.1　假设检验的一般步骤

假设检验（hypothesis testing）是先对总体的参数或分布提出假设，然后根据样本数据推断是否拒绝检验假设，检验假设也叫零假设（null hypothesis），记为 H_0，这种假设通常也被翻译为无效假设或原假设。如果从样本数据中得到的拒绝 H_0 证据不足（检验统计量 t、z、F 值对应的概率 P 值较大，如大于 0.05，不属于“小概率”），则只能不拒绝 H_0，暂且认为 H_0 成立（因为拒绝的证据不足）；如果从样本数据中得到的证据充分（检验统计量 t、z、F 值对应的概率 P 值较小，如小于或等于 0.05，属于“小概率”，即“ H_0 成立”的可能性小于或等于 5%），则拒绝 H_0，接受备择假设（alternative hypothesis，也叫研究假设（research hypothesis） H_1。

这正如法官判定一个人是否犯罪，首先是假定他“无罪”（ H_0 ），然后通过侦察寻找证据，如果证据充分则拒绝“无罪”的假定（ H_0 ），判嫌疑人有罪；否则只能暂且认为“无罪”的假定（ H_0 ）成立。

对于两样本均数比较的 t 检验，其假设检验可按下列步骤进行。

（1）建立检验假设和确定检验水准 α

零假设 H_0：$\mu_1=\mu_2$，即两样本对应的总体均数相等，差异仅由抽样误差所致。

备择假设 H_1：两样本对应的总体均数不等，其差异不仅仅是由抽样误差所致。

H_1 又分单、双侧，双侧检验（two-side test）即 H_1：$\mu_1\neq\mu_2$；单侧检验（one-side test）即 H_1：$\mu_1<\mu_2$ 或 $\mu_1>\mu_2$。具体选单侧还是双侧，需要根据专业知识确定。

检验水准是假设检验时发生第Ⅰ类错误的概率，用 α 表示，通常取 0.05，也可采用 0.1 或 0.01。

（2）计算检验统计量

根据数据类型和统计推断目的选用适当的统计量。如区间数据一般采用 t 检验、z 检验、F 检验；有序数据一般采用非参数检验；名义数据一般采用卡方检验；为了反映两个变量之间的依存关系，一般采用回归分析。

（3）确定 P 值并下结论

P 值是指在 H_0 成立的前提下，比由样本数据获得的样本检验统计量（如 z、t、F 值等）更极端的概率。P 值也是一个随机变量，由样本数据计算而得，即从同一总体中抽取的不同样本可得到不同的 P 值。若 $P\leqslant\alpha$，则拒绝 H_0，接受 H_1，即差别有统计学意义（statistically significant difference）；若 $P>\alpha$，则尚不能拒绝 H_0，差别无统计学意义。



注意： α 也是概率，它是事先人为规定的、有统计学意义的最大 P 值。

假设检验的推断结论具有概率性质，可能发生两种错误：

①H_0 正确却被拒绝，即接受错误的 H_1，称为第一类错误（或 I 型错误），第一类错误概率即检验水准 α 。

②H_0 不正确但不被拒绝，称为第二类错误（或 II 型错误），其概率用 β 表示。

8.2 样本均数与总体均数的比较

Excel 提供的两样本比较的分析工具如下

①*t* 检验：平均值的成对二样本分析；

②*F* 检验：双样本方差；

③*t* 检验：双样本等方差假设；

④*t* 检验：双样本异方差假设；

⑤*z* 检验：双样本平均差检验。在 8.2~8.4 节将逐一介绍这些工具。

借助于 Excel 函数与公式，可实现样本均数与总体均数的 *t* 检验，见光盘文件“08-01 样本与总体均数 t 检验.xlsx”。

推断样本未知总体均数 μ 与已知总体均数 μ_0（一般为理论值、标准值或大量观察的稳定值）间有无差别，可采用这一检验。

H_0：$\mu=\mu_0$

H_1：$\mu\neq\mu_0$

检验统计量：$t=\dfrac{\overline{X}-\mu_0}{S/\sqrt{n}}$，自由度 $\nu=n$-1

下面介绍两种资料类型的这类假设检验。

8.2.1 已知样本均数和标准差

例 8-1：根据大量调查，已知健康成年男子脉搏均数 μ_0 为 72 次/分。某医生在某山区抽样调查了 25 名健康成年男子，其脉搏均数为 74.2 次/分，标准差为 6.5 次/分，能否据此认为该山区健康成年男子的脉搏高于一般水平？

解：Excel 实现步骤如下：

①按图 8-1（见光盘文件“08-01 样本与总体均数 t 检验.xlsx”）键入已知数据。在单元格 C4 中键入已知总体均数 μ_0“72”，单元格 B7 中键入样本例数 *n*“25”，单元格 C7 中键入均值 $\overline{X}$“74.2”，单元格 D7 中键入标准差 *S*“6.5”。

②计算 *t* 统计量及 *P* 值等。

在单元格 G4 中键入公式“=ABS(C4-C7)/D7*B7^0.5”求 *t* 值；在单元格 G5 中键入公

式“=B7-1”求自由度；在单元格 G6 中键入公式“=TDIST(G4,G5,1)”求单侧 P 值；在单元格 G7 中键入公式“=TDIST(G4,G5,2)”求双侧 P 值；在单元格 G8 中键入公式“=C7-C4”求 $\overline{X}-\mu_0$。



	A	B	C	D	E	F	G	H	I	J	K	L
1												
2	t-检验：样本均数与总体均数的比较（已知样本均数和标准差）											
3												
4	总体均数	μ_0=	72			t值	1.69230769					
5						自由度ν	24					
6		例数n	均数$\bar{X}$	标准差S		P值（单侧）	0.052					
7	样本	25	74.2	6.5		P值（双侧）	0.104					
8						均数差 $\bar{X}-\mu_0$	2.200					

1. 已知样本均数和标准差 / 2. 已知样本原始数据

图 8-1　样本均数与总体均数比较（已知样本均数和标准差）

结果分析：根据专业知识，已知山区不会低于一般水平，故作单侧检验。t=1.692，P=0.052（单侧），按 α=0.05 水准不拒绝 H_0，可认为该山区健康成年男子的脉搏与一般人群健康成年男子的脉搏的差别无统计学意义。

温馨提示：每次只需修改该文件“1. 已知样本均数和标准差”工作表的单元格 C4、B7、C7、D7，即可得到对应的检验结果。

8.2.2　已知样本原始数据

例 8-2：正常人的脉搏均数为 72 次/分，测得 10 例某病患者的脉搏（次/分）分别为：54、67、68、78、70、66、67、70、65、69，问此病患者的脉搏次数与正常人有无差别？（见光盘文件“08-01 样本与总体均数 t 检验.xlsx”）

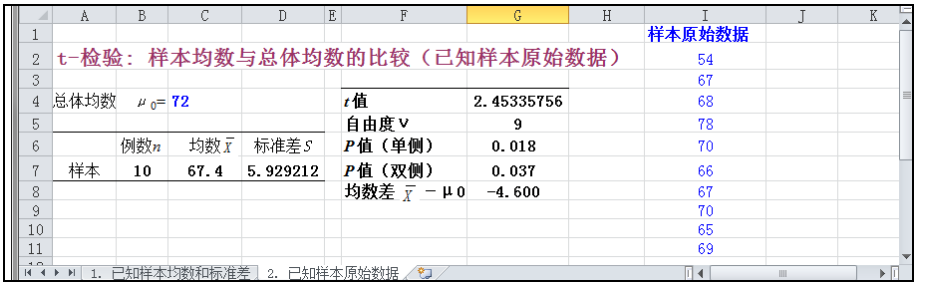
解：Excel 实现步骤如下：

① 将图 8-1 所在工作表建立工作表副本，并将工作表改名为“2. 已知样本原始数据”。

② 在单元格 I1 中键入文本“样本原始数据”，在第 I 列从单元格 I2 起依次键入 54、67、…、69 原始数据。

③ 在单元格 B7 中键入公式“=COUNT(I:I)”求原始样本数据的例数；在单元格 C7 中键入公式“=AVERAGE(I:I)”求原始样本数据的均数；在单元格 D7 中键入公式“=STDEV(I:I)”求原始样本数据的标准差。改变单元格 C4 中的总体均数 μ_0（本例仍为 72，无须改变），即可获得图 8-2。

结果分析：本例获得 t=−2.453，双侧 P=0.037，按 α=0.05 水准拒绝 H_0，接受 H_1，可认为此病患者脉搏次数与正常人的差别具有统计学意义。



	A	B	C	D	E	F	G	H	I	J	K
1									样本原始数据		
2	t-检验：样本均数与总体均数的比较（已知样本原始数据）								54		
3									67		
4	总体均数	μ_0= 72				t值	2.45335756		68		
5						自由度ν	9		78		
6		例数n	均数$\bar{X}$	标准差S		P值（单侧）	0.018		70		
7	样本	10	67.4	5.929212		P值（双侧）	0.037		66		
8						均数差 $\bar{X}-\mu_0$	-4.600		67		
9									70		
10									65		
11									69		

1. 已知样本均数和标准差　2. 已知样本原始数据

图 8-2　样本均数与总体均数比较（已知原始样本数据）

温馨提示：只需修改单元格 C4 以及第 I 列中的有关数据，即可完成样本均数与总体均数比较的 t 检验。在求例数、均数、标准差的公式中引用的单元格区域为“I:I”表示引用整个第 I 列，只要样本原始数据的例数不超过 1046576（Office 2010），本程序均有效。

8.3　配对资料的比较

随机配对设计（randomized paired design）是将受试对象按某些混杂因素（如性别、年龄、窝别等）配成对子，每对中的两个个体随机分配给两种处理组，或者同一受试对象作两次不同的处理（自身对照）。这类设计的 t 检验称为配对资料比较的 t 检验。

目的：检验同对数据之差的总体均数是否为 0。

H_0：$\mu_1-\mu_2=\mu_d=0$

H_1：$\mu_d\neq 0$

检验统计量：$t=\dfrac{\bar{d}-\mu_d}{S_{\bar{d}}}=\dfrac{\bar{d}}{S_d/\sqrt{n}}\sim t$ 分布，自由度 $\nu=n-1$，n 为对子数

例 8-3：设有 12 名志愿者服用某减肥药，服药前和服药一个疗程后各测量一次体重（kg），数据如图 8-3（也见光盘文件“08-02 配对资料的比较.xlsx”）所示。试判断此减肥药是否有效。

解：利用 Excel 提供的“t-检验：平均值的成对二样本分析”工具，可以直接进行检验。Excel 实现步骤如下：

① 按图 8-3 在工作表的单元格区域 A1:C14 中建立数据。其中单元格区域 B3:B14 为服药前数据，单元格区域 C3:C14 为服药后数据。

② 单击“数据→分析→数据分析→t-检验：平均值的成对二样本分析”，单击“确定”，弹出“t-检验：平均值的成对二样本分析”对话框，如图 8-4 所示。

	A	B	C	D	E	F	G
1	配对t检验原始数据				t-检验：成对双样本均值分析		
2	个体号	X1	X2				
3	1	101	100			X1	X2
4	2	131	136		平均	110.8333	109.5
5	3	131	126		方差	586.8788	718.6364
6	4	143	150		观测值	12	12
7	5	124	128		泊松相关系数	0.956931	
8	6	137	126		假设平均差	0	
9	7	126	116		df	11	
10	8	95	105		t Stat	0.583742	
11	9	90	87		P(T<=t) 单尾	0.285585	
12	10	67	57		t 单尾临界	1.795884	
13	11	84	74		P(T<=t) 双尾	0.571169	
14	12	101	109		t 双尾临界	2.200986	

图 8-3　配对 *t* 检验的原始数据与结果

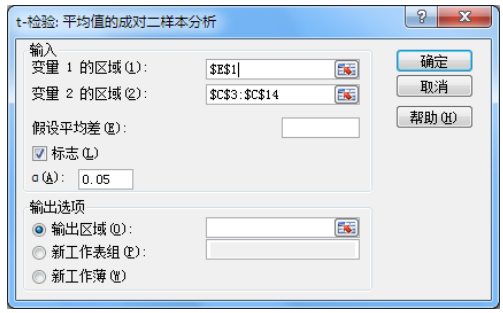


图 8-4　配对 *t* 检验对话框

③在“变量 1 的区域”中选取单元格区域 B2:B14，在“变量 2 的区域”中选取单元格区域 C2:C14，“假设平均差”键入“0”（或不键入任何值），选择“ 标志”复选框，指定检验水准 α 值（默认为 0.05），在“输出选项”中选择“输出区域”，并指定输出区域的左上角单元格地址为 E1。单击“确定”获得图 8-3 右侧的结果。

结果分析：Excel 分别输出了配对资料两个组别的一般统计量，包括平均、方差、观测值。紧接其后给出了两组的 Pearson 相关系数，指定的假设平均差（即 0）、自由度（*df*）、*t* 统计量（t Stat）、单双侧 *P* 值及 *t* 界值。本例 t=0.584，单侧 P=0.286 > 0.05，不拒绝 H_0，即尚未发现减肥药有效。

对于已知样本原始数据的样本均数与总体均数比较（见 8.2 节例 8-2），也可在相应样本原始数据的右侧，给每个个体一个总体均数。比如在图 8-2 的单元格区域 I2:I11 右侧相应单元格区域 J2:J11 中每一单元格键入“72”，然后仿照 8.3 节的配对资料比较方法，同样可获得样本均数与总体均数的比较结果。

8.4　两个样本均数的比较

两个样本均数比较的目的是推断两总体均数是否相同。即 H_0：$\mu_1 = \mu_2$，H_1：$\mu_1 \neq \mu_2$。

进行两个样本均数比较的 t 检验之前，需要判断两样本的总体方差是否齐同。两总体方差齐与不齐，所采用的 t 检验计算公式将有所不同。

下面将逐一介绍两样本的总体方差齐性（homogeneity of variance）检验、总体方差齐同的 t 检验、总体方差不齐的 t 检验和大样本的 z 检验。

8.4.1　两样本的总体方差齐性检验

若两组资料的总体方差相同，即 $\sigma_1^2 = \sigma_2^2$，则称这两个总体具有方差齐性。

H_0：$\sigma_1^2 = \sigma_2^2$

H_1：$\sigma_1^2 \neq \sigma_2^2$

检验统计量：$F=\dfrac{S_1^2}{S_2^2}$ ～F 分布，$\nu_1=n_1-1$，$\nu_2=n_2-1$

例 8-4：某克山病病区抽样测得 11 例急性克山病患者和 13 名健康人的血磷值（mmol/L）如下，问急性克山病患者和健康人的血磷值是否不同？

患者 X_1	0.84	1.05	1.20	1.20	1.39	1.53	1.67	1.80	1.87	2.07	2.11		
健康人 X_2	0.54	0.64	0.64	0.75	0.76	0.81	1.16	1.20	1.34	1.35	1.48	1.56	1.87

	A	B	C	D	E	F
1	患者X1	健康人X2		F-检验 双样本方差分析		
2	0.84	0.54				
3	1.05	0.64			患者X1	健康人X2
4	1.2	0.64		平均	1.520909091	1.084615
5	1.2	0.75		方差	0.177909091	0.17821
6	1.39	0.76		观测值	11	13
7	1.53	0.81		df	10	12
8	1.67	1.16		F	0.998310055	
9	1.8	1.2		P(F<=f) 单尾	0.506319293	
10	1.87	1.34		F 单尾临界	0.343291617	
11	2.07	1.35				
12	2.11	1.48				
13		1.56				
14		1.87				

图 8-5　总体方差齐性检验的原始数据与检验结果

解：可直接利用数据分析工具中的“*F*-检验：双样本方差”进行齐性检验。

①按图 8-5（参见光盘文件“08-03 双样本方差齐性检验.xlsx”）将观察数据键入到工作表中。

②在功能区选择“数据→分析→数据分析→*F*-检验：双样本方差”，单击“确定”。

③在弹出的“*F*-检验：双样本方差”对话框中指定参数，如图 8-6 所示。单击“确定”，获得的齐性检验结果如图 8-5 右侧所示。



图 8-6　两样本方差齐性检验对话框

结果分析：本例获得方差齐性检验 *F* 值为 0.9983，*P* 值为 0.5063>0.05，不拒绝 H_0，

即两样本的总体方差齐同。

温馨提示：一般两样本方差齐性检验 $F \geqslant 1$，P 值的 Excel 计算公式为“=FDIST(F 值, v1, v2)”。但所计算的 $F<1$ 时，表示第 2 列数据的方差大于第 1 列数据的方差，此时 P 值的 Excel 计算公式为“=1-FDIST(F 值, v1, v2)”。如例 8-4 的 $F=0.9983<1$，此时 P 值的 Excel 计算公式为“=1-FDIST(0.9983, 10, 12)”。无论两列数据谁排在前，获得的 P 值不变。

8.4.2　总体方差齐同的 t 检验

检验统计量：$t=\dfrac{\bar{X}_1-\bar{X}_2}{S_{\bar{X}_1-\bar{X}_2}} \sim t$ 分布，$\nu=n_1+n_2-2$

其中，$S_{\bar{X}_1-\bar{X}_2}=\sqrt{S_C^2\left(\dfrac{1}{n_1}+\dfrac{1}{n_2}\right)}=\sqrt{S_C^2\left(\dfrac{n_1+n_2}{n_1 n_2}\right)}$，$S_C^2=\dfrac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_1-2}$。

（1）原始数据

例 8-5：检验例 8-4 中急性克山病患者和健康人的血磷值是否不同（见光盘文件“08-04 两样本总体方差齐同的 t 检验.xlsx”）。

解：可直接使用 Excel 提供的“t-检验：双样本等方差假设”工具进行分析，具体操作如下：

①将图 8-5 中工作表单元格区域 A1:B14 观察数据复制、粘贴到一个新的工作簿中，并将此文件存为“08-04 两样本总体方差齐同的 t 检验.xlsx”。

②单击“数据→分析→数据分析→t-检验：双样本等方差假设”，单击“确定”。

③在弹出的“t-检验：双样本等方差假设”对话框中按图 8-7 所示指定参数。

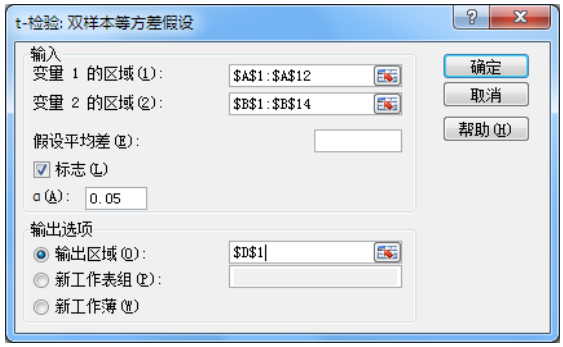


图 8-7　两样本总体方差齐同的 t 检验对话框

④单击“确定”，获得 t 检验结果如图 8-8 所示。

	A	B	C	D	E	F
1	患者X1	健康人X2		t-检验：双样本等方差假设		
2	0.84	0.54				
3	1.05	0.64			患者X1	健康人X2
4	1.2	0.64		平均	1.520909091	1.084615
5	1.2	0.75		方差	0.177909091	0.17821
6	1.39	0.76		观测值	11	13
7	1.53	0.81		合并方差	0.178073363	
8	1.67	1.16		假设平均差	0	
9	1.8	1.2		df	22	
10	1.87	1.34		t Stat	2.523724046	
11	2.07	1.35		P(T<=t) 单尾	0.00966852	
12	2.11	1.48		t 单尾临界	1.717144335	
13		1.56		P(T<=t) 双尾	0.01933704	
14		1.87		t 双尾临界	2.073873058	
15						

图 8-8　两样本总体方差齐同的 t 检验结果

结果分析：本例获得统计量 t 值为 2.5237，自由度为 22，对应的双侧 P 值为 0.0193<0.05，故拒绝 H_0，表明急性克山病患者和健康人的血磷值间的差异具有统计学意义。

（2）非原始数据

Excel 的“t-检验：双样本等方差假设”只能处理原始数据，当没有原始数据只知样本均数、标准差、例数时，可按下列方法进行 t 检验。

例 8-6：急性克山病患者和健康人血磷值的均数、标准差、例数分别见图 8-9（也见光盘文件“08-05 方差齐同的 t 检验（无原始数据）.xlsx”）的单元格区域 B3:B5、C3:C5，问急性克山病患者和健康人的血磷值是否不同？

	A	B	C	D
1	两样本无原始数据，只有均数、标准差、例数时的t检验			
2		样本1	样本2	
3	均数	1.52	1.08	
4	标准差	0.4218	0.4221	
5	例数	11	13	
6	检验水准	0.05		
7	df	22		
8	合并方差	0.178053333		
9	t Stat	2.545306115		
10	P(T<=t) 单尾	0.009221195		
11	t 单尾临界	1.717144374		
12	P(T<=t) 双尾	0.018442389		
13	t 双尾临界	2.073873068		

图 8-9　非原始数据总体方差齐同的 t 检验

解：Excel 实现按如下步骤操作。

① 按图 8-9 键入工作表中的所有文本（包括单元格区域 A3:A6，单元格 B2、C2）。在相应单元格中分别键入两样本均数、标准差和例数（见 B3:C5），在单元格 B6 中键入检验水准“0.05”，这些为已知数据。

② 计算有关统计量。在单元格 B7 中键入公式“=B5+C5–2”，计算自由度；在单元格

B8 中键入公式“=(B4^2*(B5-1)+C4^2*(C5-1))/B7”，计算合并方差；在单元格 B9 中键入公式“=(B3-C3)/SQRT(B8*(B5+C5)/(B5*C5))”，计算 *t* 检验统计量；在单元格 B10 中键入公式“=TDIST(ABS(B9),B7,1) ”，求 *t* 检验统计量对应的单侧 *P* 值；在单元格 B11 中键入公式“=TINV(B6*2,B7)”，求相应检验水准、自由度对应的 *t* 检验单侧临界值；在单元格 B12 中键入公式“=TDIST(ABS(B9),B7,2)”，求 *t* 检验统计量对应的双侧 *P* 值；在单元格　B13 中键入公式“=TINV(B6,B7)”，求相应检验水准、自由度对应的 *t* 检验双侧临界值。

结果分析：本例实际上是从例 8-5 获得的图 8-8 中摘录的已知数据，所以解释完全相同。由于小数的原因，得到的数据略有差别。

温馨提示：只需修改两样本均数、标准差和例数（即修改单元格区域 B3:C5 中的数据），并修改单元格 B6 中的“检验水准”，即可完成 *t* 检验。

8.4.3　总体方差不齐的 *t* 检验

对于两小样本均数的比较，如果方差不齐时，可以采用变量变换、秩和检验、总体方差不齐的 *t* 检验等方法进行假设检验，本节只介绍后者。

这种情况下可采用两种方法获得 *t* 临界值，一种是用校正自由度的方法查 *t* 界值表；另一种是直接用公式求出近似临界值 *t*′。Excel 采用的是第一种方法，校正自由度的计算公式为：

$$\nu = \frac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{\left(s_1^2/n_1\right)^2/(n_1-1) + \left(s_2^2/n_2\right)^2/(n_2-1)}$$

t 值计算采用：

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

利用 Excel 提供的“t-检验：双样本异方差假设”，可以直接得到检验结果。其操作步骤与前面的“t-检验：双样本等方差假设”基本一样。

例 8-7：为了与总体方差齐同的 *t* 检验结果作比较，试对例 8-4 数据进行总体方差不齐的 *t* 检验（这里仅以此例做比较，实际工作中等方差数据应进行总体方差齐同的 *t* 检验）（见光盘文件“08-06 两样本总体方差不同的 t 检验.xlsx”）。

解：直接使用 Excel 提供的“*t*-检验：双样本异方差假设”分析工具，具体操作如下：

① 将图 8-5 中工作表单元格区域 A1:B14 观察数据复制、粘贴到一个新的工作簿中，并将此文件存为“08-06 两样本总体方差不同的 t 检验.xlsx”。

② 单击“数据→分析→数据分析→*t*-检验：双样本异方差假设”，单击“确定”。

③在弹出的“*t*-检验：双样本异方差假设”对话框中按图 8-10 所示指定参数。单击“确定”，获得 *t* 检验结果如图 8-10 所示。



	A	B	C	D	E	F
1	患者X1	健康人X2		t-检验：双样本异方差假设		
2	0.84	0.54				
3	1.05	0.64			患者X1	健康人X2
4	1.2	0.64		平均	1.520909091	1.084615
5	1.2	0.75		方差	0.177909091	0.17821
6	1.39	0.76		观测值	11	13
7	1.53	0.81		假设平均差	0	
8	1.67	1.16		df	21	
9	1.8	1.2		t Stat	2.523909994	
10	1.87	1.34		P(T<=t) 单尾	0.009865233	
11	2.07	1.35		t 单尾临界	1.720742871	
12	2.11	1.48		P(T<=t) 双尾	0.019730467	
13		1.56		t 双尾临界	2.079613837	
14		1.87				

图 8-10　两样本总体方差不齐的 *t* 检验结果

结果分析：比较图 8-8 与图 8-10 可见，对于同一资料不同的 *t* 检验方法，获得的自由度、*t* 值、*P* 值等结果不同。因此在实际数据分析中，应根据数据的特征正确选择不同的检验方法。

当无原始数据，只有均数、标准差、例数时，两样本总体方差不齐的 *t* 检验计算公式如图 8-11 所示（其基本数据来自图 8-10，如样本 1 的标准差=$\sqrt{0.17791}$=0.4218，也见光盘文件“08-07 方差不齐的 t 检验（无原始数据）.xlsx”）。其获得的结果与图 8-10 相同。

	A	B	C	D	E
1	两样本无原始数据，只有均数、标准差、例数时的t检验				
2		样本1	样本2		
3	均数	1.52	1.08		
4	标准差	0.4218	0.4221		
5	例数	11	13		
6	检验水准	0.05			
7	df	22			
8	合并方差	21			
9	t Stat	2.54546378			
10	P(T<=t) 单尾	0.009414252			
11	t 单尾临界	1.720742903			
12	P(T<=t) 双尾	0.018828504			
13	t 双尾临界	2.079613845			
14					

图 8-11　两样本总体方差不齐的 *t* 检验（非原始数据）

8.4.4　大样本的 *z* 检验

当两样本例数均较多时，可近似采用 *z* 检验代替 *t* 检验。*z* 检验统计量的计算公式为：

$$z=(\bar{X}_1-\bar{X}_2)/\sqrt{S_1^2/n_1+S_2^2/n_2}$$

Excel 数据分析工具中的“z-检验：双样本平均差检验”可完成这一检验。实际上，当两样本例数均较多时，*t* 检验的结果已经十分接近于 *z* 检验的结果。*z* 检验的计算见光盘文件“08-08 z 检验.xlsx”。

对于两样本均数的比较，如果样本含量较大（如大于 100），可直接进行 z 检验。如果样本含量较小（特别是样本含量<30），则首先进行两样本方差齐性检验，方差齐同采用“双样本等方差 t 检验”，方差不齐则采用“双样本异方差 t 检验”。

此外，理论上两样本的每一组样本应分别来自正态分布总体，且两组之间相互独立。数据是否服从正态分布可采用统计学专用软件进行检验，也可采用直方图（见第 6 章）、QQ 图、PP 图等进行直观判断。

8.5　多样本方差 Bartlett 齐性检验

t 检验只能用于两样本均数的比较，当有多个样本均数进行比较时，需要采用方差分析（analysis of variance, ANOVA）。方差分析是分析处理因素方差是否与机会期望方差相等的一种方法，故此得名。该方法由英国统计学家 R.A.Fisher 于 20 世纪 20 年代首创，为纪念 Fisher，方差分析又称 F 检验。Excel 提供的方差分析工具有：单因素方差分析；无重复双因素方差分析；有重复双因素方差分析。

方差分析假定每组随机样本分别独立地来自正态分布总体，且每组总体方差齐同。考察多组方差是否齐同，可以用 Bartlett 方差齐性检验（两组样本已在 8.2 节作了 F 检验）。因为 Excel 没有提供检验多个总体方差齐性的分析工具，下面首先介绍如何用函数和公式来实现齐性检验。

H_0：各组总体方差齐同

H_1：各组总体方差不齐

统计量：$\chi^2 = \frac{1}{C}\left[(N-k)\ln S_C^2 - \sum_{i=1}^{k}(n_i-1)\ln S_i^2\right]$，自由度 $\nu = k-1$

其中，合并方差：$S_C^2 = \sum_{i=1}^{k}(n_i-1)S_i^2 \Big/ (N-k)$

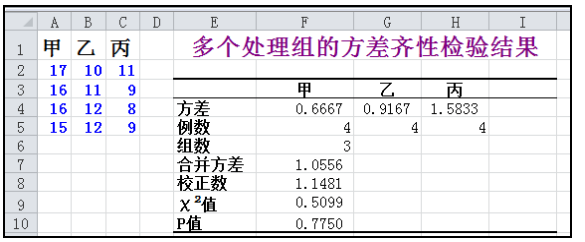
校正数：$C = 1 + \frac{1}{3(k-1)}\left[\sum_{i=1}^{k}\frac{1}{(n_i-1)} - \frac{1}{N-k}\right]$

式中，k 为组数，n_i 为各组例数，$N = \sum n_i$ 为总例数，S_i^2 为各组方差。统计量 χ^2 服从以 k-1 为自由度的 χ^2 分布。若 P 值较小，则拒绝 H_0，认为方差不齐；否则，认为方差齐同。

例 8-8：将 12 份不同个体的血液标本随机等分为 3 组，分别用甲、乙、丙 3 种抗凝剂处理后测得红细胞沉降率（mm／h）结果见图 8-12（也见光盘文件“08-09 Bartlett 齐性检验.xlsx”），问 3 种处理对应的总体方差是否齐同？

解：用 Excel 的函数和公式实现多组总体方差齐同检验的步骤如下：

①按图 8-12 键入已知条件和文本。第 A、B、C 三列为三个处理组原始数据（包括其组别名称），第 E 列为说明文本，F3:H3 为处理组名称。



	A	B	C	D	E	F	G	H	I
1	甲	乙	丙		多个处理组的方差齐性检验结果				
2	17	10	11						
3	16	11	9			甲	乙	丙	
4	16	12	8		方差	0.6667	0.9167	1.5833	
5	15	12	9		例数	4	4	4	
6					组数	3			
7					合并方差	1.0556			
8					校正数	1.1481			
9					χ^2值	0.5099			
10					P值	0.7750			

图 8-12　三个处理组的方差齐性检验

②计算有关统计量。单元格区域 F4:H5 分别为三个组的各组方差与例数，具体计算公式如下：

在单元格 F4 中键入公式“=VAR(A:A)”，计算 A 列数据的方差；将此公式向右拖放填充至单元格 H4。在单元格 F5 中键入公式“= COUNT(A:A)”，计算 A 列数据的例数；将此公式向右拖放填充至单元格 H5。

③单元格区域 F7:F10 分别为合并方差、校正数、卡方值、*P* 值等。在单元格 F7 中键入公式“=SUM(F4*(F5-1),G4*(G5-1),H4*(H5-1))/(SUM(F5:H5)-$F$6)”，计算合并方差；在单元格 F8 中键入公式“=1+(1/(3*($F$6-1)))*(SUM(1/(F5-1),1/(G5-1),1/(H5-1))-1/(SUM(F5:H5)-$F$6))”，计算校正数 *C*；在单元格 F9 中键入公式“=(1/$F$8)*((SUM(F5:H5)-$F$6)*LN($F$7)-SUM((F5-1)*LN(F4),(G5-1)*LN(G4),(H5-1)*LN(H4)))”，计算 χ^2 值；在单元格 F10 中键入公式“=CHIDIST($F$9,$F$6-1)”，计算 *P* 值。

结果分析：本例 *P*=0.775，按 α=0.05 水准，不拒绝 H_0，可认为三个处理组方差齐同。

温馨提示：按同样方法也可进行 4 个或 5 个处理组的 Bartlett 齐性检验（见光盘文件“08-09 Bartlett 齐性检验.xlsx”）。每次只需修改第 A、B、C、…列的处理组原始数据，即可获得 Bartlett 齐性检验的结果。

8.6　单因素方差分析

单因素方差分析（one-way ANOVA）主要用于完全随机设计多个独立样本均数的比较，只考虑一个研究因素，目的是考察该因素各水平处理组间的差别是否具有统计学意义。

8.6.1　单因素方差分析工具

H_0：$\mu_1=\mu_2=\cdots=\mu_k$，H_1：μ_i 不全相等。Excel 实现可采用数据分析工具“单因素方

差分析”。

例 8-9：对例 8-8 资料进行单因素方差分析。

解：单因素方差分析步骤如下：

①按图 8-13 在工作表的单元格区域 A1:C5 中建立数据。

	A	B	C	D	E	F	G	H	I	J	K
1	甲	乙	丙		方差分析：单因素方差分析						
2	17	10	11								
3	16	11	9		SUMMARY						
4	16	12	8		组	观测数	求和	平均	方差		
5	15	12	9		甲	4	64	16	0.666667		
6					乙	4	45	11.25	0.916667		
7					丙	4	37	9.25	1.583333		
8											
9											
10					方差分析						
11					差异源	SS	df	MS	F	P-value	F crit
12					组间	96.1667	2	48.08333	45.55263	1.96E-05	4.256495
13					组内	9.5	9	1.055556			
14											
15					总计	105.667	11				

图 8-13　单因素方差分析结果

②单击“数据→分析→数据分析→方差分析：单因素方差分析”，单击“确定”，弹出“方差分析：单因素方差分析”对话框，如图 8-14 所示。

③如图 8-14 所示填入参数，单击“确定”后可获得图 8-13 中的结果。

结果分析：Excel 输出两张表格，第一张表汇总（SUMMARY），给出了参加比较的各处理组例数、求和、均数和方差；第二张表输出了方差分析结果，其计算的数学公式见表 8-1。本例得到统计量 F=45.55263，P=1.96×10^{-5}<0.05，拒绝 H_0，认为 3 个处理组总体均数的差别具有统计学意义。（结果中“F crit”是检验水准为 0.05 时的 F 界值。）



图 8-14　方差分析：“单因素方差分析”对话框

表中 k 为组数，n_i 为各组例数，$N=\sum n_i$ 为总例数，X_{ij} 为第 i 组第 j 个体的值，$\bar{X}_i$ 为第 i 组的均数，$\bar{X}$ 为总均数。

表 8-1　单因素方差分析的计算表

变异来源	离均差平方和 SS	自由度 v	均方 MS	F 值
组间	$\sum_{i=1}^{k} n_i(\bar{X}_i-\bar{X})^2$	$k-1$	$\frac{SS_{组间}}{v_{组间}}$	$\frac{MS_{组间}}{MS_{组内}}$
组内	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_i)^2$	$N-k$	$\frac{SS_{组内}}{v_{组内}}$	
总计	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\bar{X})^2$	$N-1$		

Bartlett 检验要求资料服从正态分布，当数据的分布明显非正态时，其偏差较大；故近年来采用更多的是 Levene 检验，该法不依赖于总体分布的具体形式。Levene 检验法既可用于两个总体方差齐性检验，也可用于多个总体方差齐性检验，所分析的资料可不具正态性。该法是将 $Z_{ij}=\left|X_{ij}-\bar{X}_i\right|$ 当作测量数据进行上面所介绍的单因素方差分析。其中 X_{ij} 为原始观测值，$\bar{X}_i$ 为各处理组样本均数。对于例 8-8 数据，计算得到的 Z_{ij} 见光盘文件“08-10 单因素方差分析.xlsx”的“Levene 方差齐性检验”工作表的单元格区域 F3:H6。对该区域数据利用 Excel 工具“方差分析：单因素方差分析”（参见本节前面叙述）得到 F 检验 P 值为 0.6743，结果与 Bartlett 检验 P＝0.7750 很接近，均不拒绝 H_0，结论是该数据方差齐同。

8.6.2　单因素方差分析图示

光盘文件“08-11 单因素方差分析图示.xlsx”有 3 个工作表，各分别可显示 3、4、5 个处理组的散点及均数的情况，可大致看出各组数据之间的差别。图 8-15 为含有 3 个处理组的图形，遇到类似数据，可将第 1~3 组数据分别存放在第 A～C 列中。程序会自动计算绘图所需要的数据 E3:G12 与第 K 列至第 M 列。

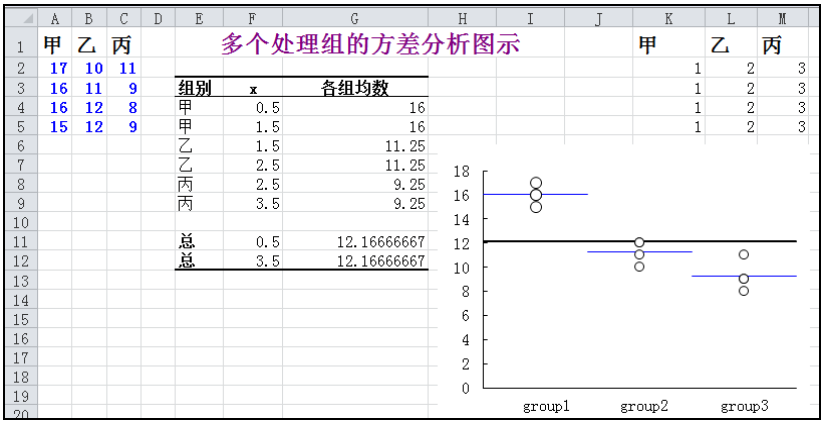


图 8-15　三组数据图示

结果分析：第 1 组数据明显高于第 2、3 组，位于平均水平（见中间粗的长水平线）的上方。第 2、3 组位于平均水平下方，差别不十分明显。

温馨提示：将已知各组数据分别键入或复制、粘贴到相应列（限制在 1000 行以内）。在图形上的散点位置单击鼠标右键，根据各组数据的例数修改各组（group1、group2、group3）源数据系列的 X 值单元格范围（因为此处尚不能自动修改，需要人工修改）。光盘文件“08-11 单因素方差分析图示.xlsx”有 3 个工作表，各分别可显示 3、4、5 个处理组。对于光盘中的“4 组”或“5 组”工作表也可进行类似修改。

8.6.3　无原始数据的单因素方差分析

当阅读科技文献或进行稿件审核时，往往没有原始数据，为了考证原作者的假设检验是否正确，可采用光盘文件“08-12 无原始数据的单因素方差分析.xlsx”。这里计算总平均时采用加权平均，即 $\bar{X}=\sum n_i\bar{X}_i/\sum n_i$，其他计算见表 8-1 中公式。

结果分析：以上数据实际上是例 8-8、例 8-9 的计算结果，所以图 8-16 显示的结果亦与图 8-13 相同。此外，在图 8-16 的右下方还显示了这三组数据的均数与标准差。

温馨提示：该程序可比较 20 个处理组，将例数、均数、标准差分别键入或复制、粘贴到相应的 B、C、D 列，可获得单因素方差分析的结果。

8.6.4　单因素方差分析的多重比较

若方差分析的结论是拒绝无效假设 H_0，接受 H_1，只能认为多个总体均数不全相等。如果想了解究竟哪几个均数间存在差异，需要进行多个样本均数间的两两比较，两两比较也称为事后多重比较（Post hoc multiple comparisons）。样本均数间的两两比较不能用两样本均数比较的 t 检验。原因有两点：

①增加了假设检验的次数，如有 k 个分组，即 k 水平进行 k 个均数的比较需要进行

$$c=C_k^2=k(k-1)/2$$

次 t 检验，如 k=3，4，5，6，…，则分别需要进行 3，6，10，15，…次 t 检验。

②增大了 I 型错误的累计概率。根据二项分布原理，在检验水准 α 情况下，如果进行 c 次 t 检验，则至少有一次犯 I 型错误的概率为：

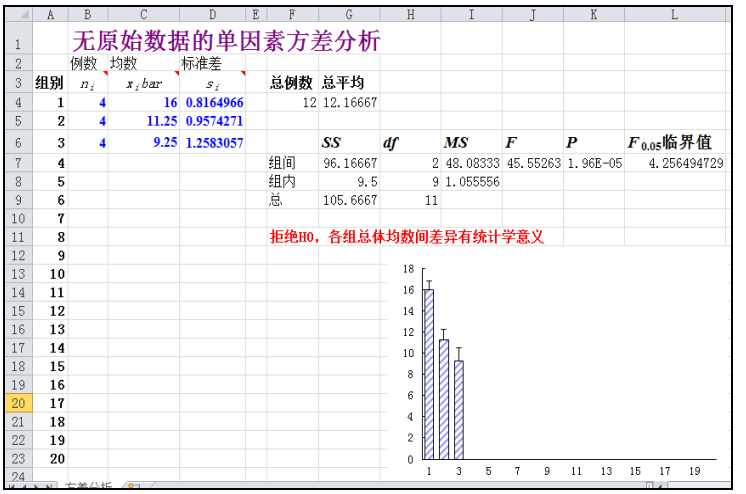
$$\alpha_E=P(X\geqslant 1)=1-P(X=0)=1-C_c^0\times\alpha^0(1-\alpha)^c=1-(1-\alpha)^c$$

在检验水准 α =0.05 情况下，如果进行 15 次 t 检验，则至少有一次犯 I 型错误的概率为：

$$P(X\geqslant 1)=1-(1-0.05)^{15}=0.5367$$

其中，后一点是不能利用一般 t 检验的关键原因。

下面介绍两种常用的多重比较方法。



无原始数据的单因素方差分析

组别	例数 n_i	均数 $\bar{x}_i$	标准差 s_i
1	4	16	0.8164966
2	4	11.25	0.9574271
3	4	9.25	1.2583057
4			
5			
6			
7			
8			
9			
10			
11			
12			
13			
14			
15			
16			
17			
18			
19			
20			

总例数	总平均
12	12.16667

	SS	df	MS	F	P	$F_{0.05}$临界值
组间	96.16667	2	48.08333	45.55263	1.96E-05	4.256494729
组内	9.5	9	1.055556			
总	105.6667	11				

拒绝H0，各组总体均数间差异有统计学意义

图 8-16　无原始数据的单因素方差分析

1. LSD-*t* 检验

这是最小有意义差异（least significant difference）*t* 检验，是基于两样本 *t* 检验原理进行的检验，回忆 8.4.2 节检验统计量：

$$t=\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{S_{\bar{X}_1-\bar{X}_2}}=\frac{\bar{X}_1-\bar{X}_2}{\sqrt{S_C^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}},\quad \nu=n_1+n_2-2$$

而 LSD-*t* 检验的 *t* 值计算公式为：

$$t=\frac{\left(\bar{X}_A-\bar{X}_B\right)-(\mu_A-\mu_B)}{S_{\bar{X}_A-\bar{X}_B}}=\frac{\bar{X}_A-\bar{X}_B}{\sqrt{MS_{\text{误差}}\left(\frac{1}{n_A}+\frac{1}{n_B}\right)}},\quad \nu=N-k$$

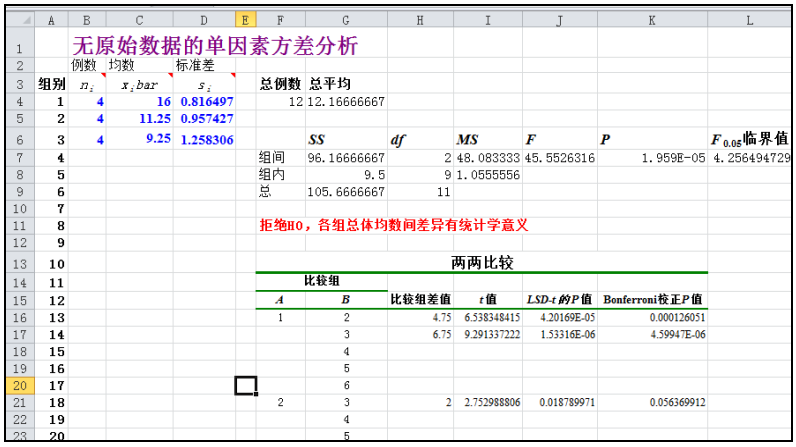
$\bar{X}_A$、$\bar{X}_B$ 分别为两比较组的样本均数，$MS_{\text{误差}}$ 为方差分析表中的 $MS_{\text{组内}}$，n_A、n_B 分别为两比较组的样本含量，N 为各组的样本含量之和，k 为分组组数，ν 为方差分析中的组内自由度。

2. Bonferroni 法

如果进行 c 次 *t* 检验，则至少有一次犯 I 型错误的概率为 $\alpha_E=1-(1-\alpha)^c$，进行多次两两 *t* 检验将会大大增加得到犯 I 型错误的概率。Bonferroni 法考虑了累计犯 I 型错误的概率问题，是在 LSD-*t* 检验基础上对检验水准进行适当调整的一种保守方法（更不容易拒绝 H_0）。该方法令 $\alpha_E=c\alpha$，而不是 $\alpha_E=1-(1-\alpha)^c$。Bonferroni 校正 *P* 值为 LSD-t 检验 *P* 值

的 c 倍。Bonferroni 法是两两比较方法中最为保守的方法，当比较次数 c 不多时，该法效果较好；当比较次数 c 较多时（如 c>10），由于其检验水平选择较低，结论偏于保守。

光盘文件“08-13 单因素方差分析的多重比较.xlsx”根据以上原理制成，其界面如图 8-17 所示。



	A	B	C	D	E	F	G	H	I	J	K	L
1		无原始数据的单因素方差分析										
2		例数	均数	标准差								
3	组别	n_i	$x_i bar$	s_i		总例数	总平均					
4	1	4	16	0.816497		12	12.16666667					
5	2	4	11.25	0.957427								
6	3	4	9.25	1.258306			SS	df	MS	F	P	$F_{0.05}$临界值
7	4					组间	96.16666667	2	48.083333	45.5526316	1.959E-05	4.256494729
8	5					组内	9.5	9	1.0555556			
9	6					总	105.6666667	11				
10	7											
11	8					拒绝H0，各组总体均数间差异有统计学意义						
12	9											
13	10						两两比较					
14	11						比较组					
15	12						A	B	比较组差值	t值	LSD-t的P值	Bonferroni校正P值
16	13						1	2	4.75	6.538348415	4.20169E-05	0.000126051
17	14							3	6.75	9.291337222	1.53316E-06	4.5994TE-06
18	15							4				
19	16							5				
20	17							6				
21	18						2	3	2	2.752988806	0.018789971	0.056369912
22	19							4				
23	20							5				

图 8-17　单因素方差分析的多重比较

结果分析：图 8-17 除显示了与图 8-16 相同的方差分析表外，下方还显示了这三组数据的多重比较结果。如按 Bonferroni 结果判断，在 $\alpha=0.05$ 检验水准下，第 2 组与第 3 组数据差异无统计学意义，其他有统计学意义。该结果也与图 8-15 相吻合。

温馨提示：该程序可进行 6 个处理组以内的两两比较，将例数、均数、标准差分别键入或复制、粘贴到相应的 B、C、D 列，可获得单因素方差分析和相应的多重比较结果。

8.7　双因素方差分析

如果按两个因素纵横排列，两个因素不同水平交叉对应格子仅有一个值时，称这种数据为无重复数据，应采用无重复双因素方差分析。这一方法是配对设计资料 t 检验的扩展，所以也叫随机配伍组设计的方差分析。如果两个因素不同水平交叉对应格子不仅仅只有一个值，而有两个或两个以上值时，称为有重复数据，应采用有重复双因素方差分析，后者可考虑因素间的交互效应，有的教科书也称这种分析为两因素析因分析。

8.7.1　无重复双因素方差分析

无重复双因素方差分析通常也称为随机配伍组设计或随机区组设计的方差分析。这种

数据与单因素方差分析数据的明显区别在于：研究者不仅仅对处理组不同水平之间有无差别感兴趣，而且也对区组个体间的差别感兴趣，每一个处理组一定有相同的个体数，且每一区组个体有明确的编号。而单因素方差分析数据的各处理组例数可以相等也可以不相等，个体通常无明确的编号。

H_{01}：各区组总体均数相同；　　H_{11}：各区组总体均数不全相同。

H_{02}：各处理组总体均数相同；　　H_{12}：各处理组总体均数不全相同。

随机区组设计方差分析将总变异分解为处理组间变异、区组间变异和误差三个部分，将自由度也作相应的分解，方差分析计算表见表 8-2。

表 8-2　无重复双因素方差分析计算表

变异来源	*SS*	ν	*MS*	*F*
区组间	$SS_{B_1}=\sum_{j=1}^{n}k(\bar{X}_j-\bar{X})^2$	$\nu_{B_1}=n-1$	SS_{B_1}/ν_{B_1}	MS_{B_1}/MS_E
处理间	$SS_{B_2}=\sum_{i=1}^{k}n(\bar{X}_i-\bar{X})^2$	$\nu_{B_2}=k-1$	SS_{B_2}/ν_{B_2}	MS_{B_2}/MS_E
误差	$SS_E=SS_T-SS_{B_1}-SS_{B_2}$	$\nu_E=\nu_T-\nu_{B_1}-\nu_{B_2}$	SS_E/ν_E	
总计	$SS_T=\sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij}-\bar{X})^2$	$\nu_T=kn$-1		

其中 k 为处理组数，n 为区组数，X_{ij} 为第 i 处理组第 j 区组的值，$\bar{X}_i$ 为第 i 处理组的均数，$\bar{X}_j$ 为第 j 区组的均数，$\bar{X}$ 为总均数。

例 8-10：三批甘蓝叶样本分别在甲、乙、丙、丁 4 种条件下测量核黄素浓度，试验结果见图 8-18（也见光盘文件“08-14 无重复双因素方差分析.xlsx”）左侧所示。问 4 种条件下的测量结果的差异是否具有统计学意义？

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		测量条件					方差分析：无重复双因素分析						
2	批次	甲	乙	丙	丁								
3	1	27.2	24.6	39.5	38.6		SUMMARY	观测数	求和	平均	方差		
4	2	23.2	24.2	43.1	39.5		1	4	129.9	32.475	58.9025		
5	3	24.8	22.2	45.2	33		2	4	130	32.5	105.58		
6							3	4	125.2	31.3	107.0533		
7													
8							甲	3	75.2	25.06667	4.053333		
9							乙	3	71	23.66667	1.653333		
10							丙	3	127.8	42.6	8.31		
11							丁	3	111.1	37.03333	12.40333		
12													
13													
14							方差分析						
15							差异源	SS	df	MS	F	P-value	F crit
16							行	3.761667	2	1.880833	0.229939	0.801274	5.143253
17							列	765.5292	3	255.1764	31.19622	0.000467	4.757063
18							误差	49.07833	6	8.179722			
19													
20							总计	818.3692	11				

图 8-18　无重复双因素方差分析结果

解：无重复双因素方差分析可按下列步骤进行。

①单击“数据工具→数据分析→方差分析：无重复双因素分析”，单击“确定”，弹出“方差分析：无重复双因素分析”对话框，如图 8-19 所示。

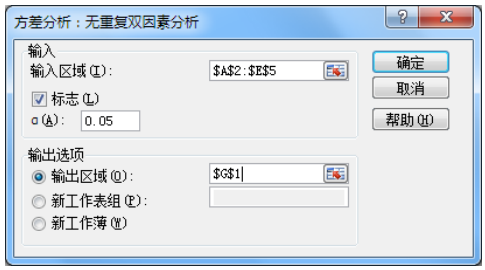


图 8-19 “方差分析：无重复双因素分析”对话框

②如图 8-19 所示填入参数（选取单元格区域 A2:E5），单击“确定”可获得图 8-18 右侧的结果。

结果分析：结果分为两个部分：前一部分为汇总（SUMMARY），分别按行和列给出了各配伍组与处理组的例数、求和、均数和方差；后一部分为方差分析结果，包括行因素（区组因素）和列因素（处理组因素）的 F 值、P 值和 F 临界值。本例区组间（行）F=0.2299，P=0.80127>0.05，处理组间（列）F=31.1962，P=0.000467<0.05，故认为 4 种条件下测量结果的差异具有统计学意义，而各批次间差异无统计学意义。

8.7.2　有重复双因素方差分析

在科学研究中，许多时候两因素的联合效应不等于两因素单独效应之和，存在拮抗作用或协同作用，此时就要考察其交互作用，相应的方差分析称为析因设计的方差分析。因为要检验交互作用的效应是否有意义，对于两个因素的每一组合至少要做 2 次试验（重复），否则就不能将交互作用与误差分离开来，故又称为有重复双因素方差分析。

该方法要求各组样本重复次数相等，可以执行完全随机设计的析因分析。

假设两个因素分别为 A 因素与 B 因素。总变异分解为处理组间变异和误差变异，总自由度也相应分解为处理组间自由度和误差自由度；再将处理组间的变异与自由度分解成 A 因素、B 因素及 AB 交互作用三个部分，分别用各部分均方与误差均方作比较，从而考察 A、B 因素的主效应及交互作用的影响，如表 8-3 所示。

其中 I 为 A 因素水平数，A_i（i=1,2,⋯,I）为 A 因素各水平的合计；J 为 B 因素的水平数，B_j（j=1,2,⋯,J）为 B 因素各水平的合计；T_k（k=1,2,⋯,G）为各处理组合计；r 为每组例数，即重复数；$G=IJ$ 为总处理组数。$C=\left(\sum X\right)^2$。

表 8-3　有重复双因素方差分析计算表

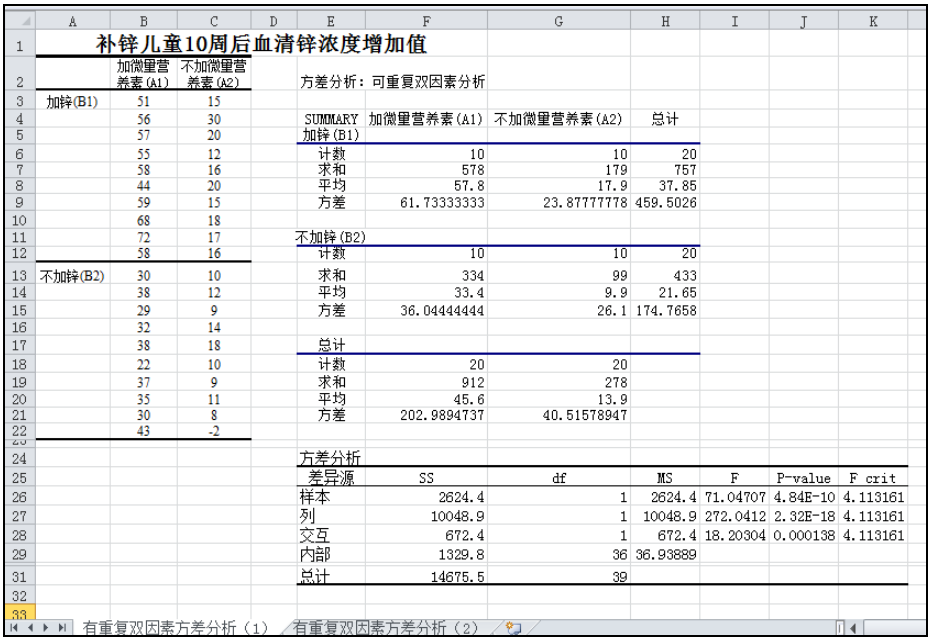
来　源	SS	v	MS	F
处理组间	$SS_B=\frac{1}{r}\sum T_i^2-C$	$v_B=G-1$		
A	$SS_{(A)}=\frac{1}{rJ}\sum A_i^2-C$	$v_{(A)}=I-1$	$SS_{(A)}/v_{(A)}$	$MS_{(A)}/MS_E$
B	$SS_{(B)}=\frac{1}{rI}\sum B_j^2-C$	$v_{(B)}=J-1$	$SS_{(B)}/v_{(B)}$	$MS_{(B)}/MS_E$
AB	$SS_{(AB)}=SS_B-SS_{(A)}-SS_{(B)}$	$v_{(AB)}=$(I-1)(J-1)	$SS_{(AB)}/v_{(AB)}$	$MS_{(AB)}/MS_E$
误差（内部）	$SS_E=SS_T-SS_B$	$v_E=v_T-v_B$	SS_E/v_E	
总计	$SS_T=\sum X^2-C$	$v_T=rG$-1		

H_{01}：A 因素各组总体均数相同；　H_{11}：A 因素各组总体均数不全相同。

H_{02}：B 因素各组总体均数相同；　H_{12}：B 因素各组总体均数不全相同。

H_{03}：A、B 因素无交互作用；H_{13}：A、B 因素有交互作用。

例 8-11：图 8-20（也见光盘文件“08-15 有重复双因素方差分析.xlsx”）的单元格区域 A1:C22 中的数据是 40 名缺锌儿童用不同补锌方法治疗 10 周后血清锌浓度的增加值（μg/100ml），试分析微量营养素、锌及微量营养素与锌的共同作用。



补锌儿童10周后血清锌浓度增加值

	加微量营养素(A1)	不加微量营养素(A2)
加锌(B1)	51	15
	56	30
	57	20
	55	12
	58	16
	44	20
	59	15
	68	18
	72	17
	58	16
不加锌(B2)	30	10
	38	12
	29	9
	32	14
	38	18
	22	10
	37	9
	35	11
	30	8
	43	-2

方差分析：可重复双因素分析

SUMMARY	加微量营养素(A1)	不加微量营养素(A2)	总计
加锌(B1)			
计数	10	10	20
求和	578	179	757
平均	57.8	17.9	37.85
方差	61.73333333	23.87777778	459.5026
不加锌(B2)			
计数	10	10	20
求和	334	99	433
平均	33.4	9.9	21.65
方差	36.04444444	26.1	174.7658
总计			
计数	20	20	
求和	912	278	
平均	45.6	13.9	
方差	202.9894737	40.51578947	

方差分析

差异源	SS	df	MS	F	P-value	F crit
样本	2624.4	1	2624.4	71.04707	4.84E-10	4.113161
列	10048.9	1	10048.9	272.0412	2.32E-18	4.113161
交互	672.4	1	672.4	18.20304	0.000138	4.113161
内部	1329.8	36	36.93889			
总计	14675.5	39				

图 8-20　40 名缺锌儿童补锌治疗结果及部分方差分析结果数据

解：有重复双因素方差分析可按下列步骤进行。

① 按图 8-20 在工作表的单元格区域 A1:C22 中键入数据。请注意数据存放格式，否则

无法完成，或获得错误结果。

② 单击“数据→分析→数据分析→方差分析：可重复双因素分析”，单击“确定”，弹出“方差分析：可重复双因素分析”对话框，如图 8-21 所示。

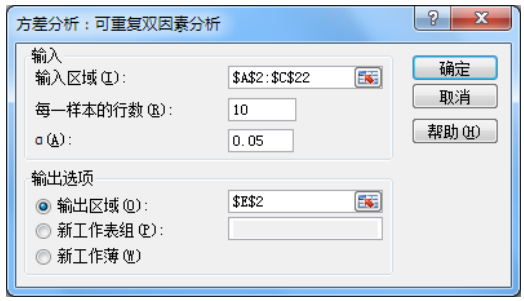


图 8-21　可重复双因素方差分析对话框

③ 按图 8-21 所示填入参数（选取单元格区域 A2:C22），单击“确定”可获得图 8-20 右侧结果。

结果分析：结果分为两个部分：前一部分为汇总（SUMMARY），有三个表，第一、二张表汇总，分别给出按行和列分组后各组例数、求和、均数和方差；第三张表给出了按 A 因素分组（忽略 B 因素）后的合计。最后一部分为方差分析结果，包括样本（行因素 A）、列因素（因素 B）和交互作用（AB）的 F 值、P 值和 F 临界值。

从方差分析结果可知，微量营养素、锌及交互作用均 $P<0.001$，即微量营养素和锌对血清锌浓度的影响有统计学意义，且二者存在交互作用。从各组平均值来看，加微量营养素高于不加微量营养素，加锌高于不加锌，同时加微量营养素和锌血清锌浓度最高。

在光盘文件的另一个工作表中，可见“20 只家兔神经缝合后的轴突通过率（%）”的例子，读者不妨对此数据进行有重复双因素方差分析。读者应特别注意数据的存放格式，否则可能得不到所需结果。

第 9 章

名义与有序数据的假设检验

名义数据通常采用 χ^2 检验，有序数据通常采用非参数检验，本章将依次介绍 Excel 实现这些统计分析的方法。

9.1　卡方检验

在第 5 章的 5.2.4 节，两个名义变量经整理便成了二维列联表，对于这类数据通常可计算率或构成比，对于率或构成比的假设检验通常采用 χ^2 检验。因为 Excel 没有提供现成的 χ^2 检验分析工具，故本章主要介绍如何利用 Excel 的公式和函数来实现 χ^2 检验。

9.1.1　四格表资料的 χ^2 检验

四格表也被称为 2×2 列联表，原始数据有 a、b、c、d 四个格子（见表 9-1），故此而得名。四格表资料的 χ^2 检验用于两个率（或构成比）的差别比较。

例 9-1：为了解铅中毒病人是否有尿棕色素增加，分别对病例组和对照组的尿液作尿棕色素定性检查，结果见表 9-1。问铅中毒病例与对照人群的尿棕色素阳性率有无差别？

表 9-1　两组人群尿棕色素阳性率比较

组　别	阳 性 数	阴 性 数	合　计	阳性率（%）
铅中毒病例	29（a）	7（b）	36（$a+b$）	80.56
对照组	9（c）	28（d）	37（$c+d$）	24.32
合计	38（$a+c$）	35（$b+d$）	73（n）	52.05

解：H_0：两总体率相等，$\pi_1=\pi_2$

H_1：$\pi_1\neq\pi_2$ $\alpha=0.05$

此时统计量可以用卡方检验的基本公式：

$$\chi^2=\sum\frac{(A-T)^2}{T}，\ \nu=（行数-1）\times（列数-1）$$

或

$$\chi^2=\sum\frac{(O-E)^2}{E}$$

式中，A 是实际频数（actual frequency，如表 9-1 中的 a=29、b=7、c=9、d=28）或观察频数（observed frequency）；T 是理论频数（theoreticalfrequency）或期望频数（expected frequency），为所在行、列合计的乘积除以总例数，即 $T=\dfrac{行合计\times列合计}{总例数}=\dfrac{n_R n_C}{n}$，$n=a+b+c+d$，$n_{1.}=a+b$、$n_{2.}=c+d$、$n_{.1}=a+c$、$n_{.2}=b+d$。如实际频数 29 的理论频数 T=（36×38/73）=18.74，其他分别为 17.26、19.26、17.74。光盘文件“02-02 四格表卡方检验理论频数.xlsx”可以方便计算理论频数。四格表为 2 行 2 列，所以自由度 ν=（2–1）×（2–1）

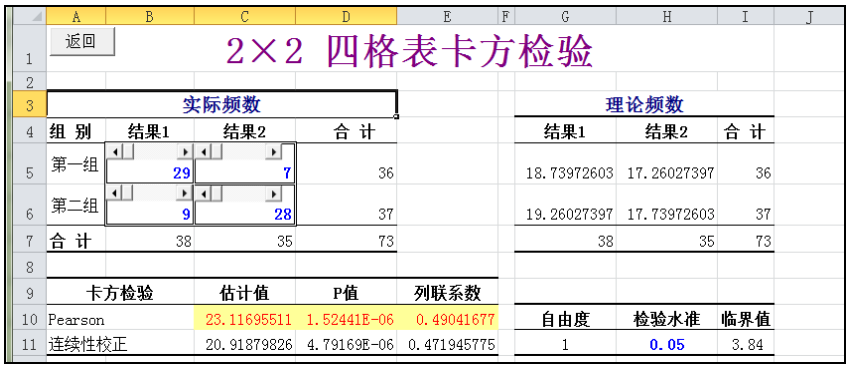
=1。

以上卡方检验基本公式的统计分析结果通常也称之为 Pearson 卡方（Pearson Chi-square）检验结果。Pearson 卡方检验要求总例数 $n \geqslant 40$，且每个格子理论频数 $T \geqslant 5$。如果 $n \geqslant 40$，但有至少 1 个格子理论频数 $1 \leqslant T < 5$，则需要采用连续性校正（continuity correction）公式：

$$\chi^2 = \sum \frac{(|A-T|-0.5)^2}{T}，\ \nu =（行数-1）\times（列数-1）$$

如果 $T<1$ 或 $n<40$ 时，则一般需采用四格表资料的 Fisher 精确概率法。

本例 $n=73>40$，理论频数 T 均大于 5，所以采用 Pearson 卡方检验结果。



返回

2×2 四格表卡方检验

实际频数			
组 别	结果1	结果2	合 计
第一组	29	7	36
第二组	9	28	37
合 计	38	35	73

理论频数		
结果1	结果2	合 计
18.73972603	17.26027397	36
19.26027397	17.73972603	37
38	35	73

卡方检验	估计值	P值	列联系数
Pearson	23.11695511	1.52441E-06	0.49041677
连续性校正	20.91879826	4.79169E-06	0.471945775

自由度	检验水准	临界值
1	0.05	3.84

图 9-1　四格表资料的卡方检验

图 9-1 所显示的工作表共分为 5 个部分：

①第一行为整个检验的标题；

②实际频数；

③理论频数；

④卡方检验结果；

⑤检验临界值。

以上四格表资料的卡方检验可按如下步骤进行。

① 按图 9-1（也见光盘文件“09-01 卡方检验.xlsm”的“2×2 表资料”工作表，注意“xlsm”后缀文件表示工作簿中含有“宏”，打开这类文件一般可直接启用宏，如果宏被禁止请启用宏）在单元格区域 A3:D7 中键入文本和已知的四格表数据。在单元格 D5 中键入公式“=B5+C5”，并将此公式拖放到单元格 D6，求行合计；在单元格 B7 中键入公式“=B5+B6”，并将此公式拖放到 D7，分别求列合计与总例数。

② 选取复制单元格区域 B3:D7，粘贴到 G3。修改“实际频数”为“理论频数”；选取单元格区域 G5:I7，清除数据内容。在单元格 I5 中键入公式“=D5”，将此公式复制、粘贴到其他行、列合计与总例数格子中。在单元格 G5 中键入公式“=$I5*G$7/$I$7”，求四格表第一个

格子的理论频数，将该公式拖放填充至四格表的其他 3 个格子，求其他格子的理论频数。

③ 根据卡方检验基本公式，在单元格 C10 中键入“=SUM((ABS(B5-G5))^2/G5,(ABS(C5-H5))^2/H5,(ABS(B6-G6))^2/G6,(ABS(C6-H6))^2/H6)”，计算 Pearson 卡方值；在单元格 C11 中键入公式“=SUM((ABS(B5-G5)-0.5)^2/G5,(ABS(C5-H5)-0.5)^2/H5,(ABS(B6-G6)-0.5)^2/G6,(ABS(C6-H6)-0.5)^2/H6)”，计算连续性校正卡方值。单元格 E10、E11 求列联系数。

④ 在单元格 G11 中键入自由度为“1”，在 H11 中键入检验水准为“0.05”，在 I11 中键入公式“=CHIINV(H11,G11)”求临界值。

⑤ 在单元格 D10 中键入公式“=CHIDIST(C10,$G$11)”，计算 Pearson 卡方值对应的 P 值；拖放该公式到单元格 D11 中，计算连续性校正卡方值对应的 P 值。

⑥ 为了使用方便，在实际频数格子中采用单击“开发工具→控件→插入”制作了滚动条（工作表已被保护，密码为“0”，去掉保护方可进行滚动条的制作），读者使用时既可以直接键入数据，也可以滑动滚动条选取数据。程序采用“条件格式”显示结果。

结果分析： 本例 n>40，理论频数都大于 5，故选用 Pearson 卡方检验结果，获得 χ^2=23.117，P=1.524×10^{-6}，按检验水准 α=0.05 拒绝 H_0，认为两组尿棕色素阳性率差别有统计学意义。



图 9-2　卡方检验程序选择界面

温馨提示： 为了读者使用方便，特将所有卡方检验（最大行数与列数均为 5）集成到了一个文件，即“09-01 卡方检验.xlsm”。打开文件后弹出图 9-2 界面，根据数据排列的行数与列数，来选定恰当的“行×列表资料”，如四格表资料选择“2×2 表资料”，见图 9-2 所示。

如果在如图 9-2 所示的界面中暂不能进入某一选项，请更换一下资料类型选择其他选项，“返回”后，再次单击所需选项即可进入其中的所需工作表。

上述例子选择“1.2×2 表资料”，进入如图 9-1 所示的界面，可采用滑动滚动条或直接键入数据方式，修改四格表实际频数（呈蓝色字体，本书一般情况下以蓝色字体所在单元格作为可修改单元格）便可得到所需要的卡方检验结果。如果还想获得除 0.05 以外的临界值，还可以修改“检验水准”下方的数字（呈蓝色字体）。读取结果时注意理论频数的提示，如果出现红色提示，则注意是否应选择连续性校正卡方值及其对应 P 值，而不选择 Pearson 卡方值及其对应 P 值。单击左上角的“返回”按钮，可回到如图 9-2 所示的界面。下面来看一个理论频数小于 5 的例子。

例 9-2：为评价中西结合治疗抑郁发作的疗效，将 155 例患者随机分为两组，两组患者均选用阿咪替林或氟西汀等西医综合治疗，中西医结合组在上述治疗的同时，再配合中医辨证治疗，根据中医辨证分型采用不同的方剂，结果见表 9-2。问两组的疗效有无差别？

表 9-2　中西结合治疗抑郁发作的疗效

组　别	有　效	无　效	合　计
中西医结合组	72	2	74
西医组	75	6	81
合计	147	8	155

将表 9-2 中的四格表数据键入到“09-01 卡方检验.xlsm”的“1.2×2 表资料”工作表后，可得到图 9-3。由“理论频数”部分所显示的结果可见，尽管总例数 n=155>40，但有四格表 c、d 格子的理论频数小于 5（注意理论频数下方有红色字提示“警告：至少有一个格子的理论频数小于 5”），所以本例应采用连续性校正卡方值 $\chi^2=0.920$，对应的 P=0.338，列联系数为 0.0768。所以由此结果可见，样本资料尚不能说明两组疗效差异有统计学意义。

返回

2×2 四格表卡方检验

实际频数			
组 别	结果1	结果2	合 计
第一组	72	2	74
第二组	75	6	81
合 计	147	8	155

理论频数		
结果1	结果2	合 计
70.18064516	3.819354839	74
76.81935484	4.180645161	81
147	8	155

警告：至少有一个格子的理论频数小于5

卡方检验	估计值	P值	列联系数
Pearson	1.748661927	0.186045035	0.105621178
连续性校正	0.919590048	0.337582635	0.076797448

自由度	检验水准	临界值
1	0.05	3.84

图 9-3　连续性校正卡方检验

9.1.2　四格表配对资料的χ^2检验

与 8.2 节配对资料的 t 检验数据类似，都是将受试对象按某些混杂因素（如性别、年龄、窝别等）配成对子，每对中的两个个体随机分配给两种处理之中，或者同一受试对象作两次不同的处理（自身对照），然后观测结果。其与 t 检验数据的差别在于结果变量不再是区间连续变量，而是名义二分类变量，如阳性、阴性，有病、无病，生存、死亡等。

例 9-3：某研究室用甲、乙两种血清学方法检查鼻咽癌患者血清 93 份，结果两法都是阳性的 45 份，都是阴性的 20 份；甲法阳性但乙法阴性的 22 份，甲法阴性而乙法阳性的 6 份。问两法检出率有无差别？

表 9-3　配对设计 2×2 列联表形式

甲　法	乙　法		合　计
	+	–	
+	45（a）	22（b）	$a+b$
–	6（c）	20（d）	$c+d$
合计	$a+c$	$b+d$	$n=a+b+c+d$

返回

2×2表配对资料卡方检验

实际频数

甲法 \ 乙法	+	–	合 计
+	45	22	67
—	6	20	26
	51	42	93

警告：(b+c)=28<40

检验结果

卡方检验	估计值	P值
非校正	9.14285714	0.002496909
连续性校正	8.03571429	0.004586392

自由度	检验水准	临界值
1	0.05	3.84

二项分布精确检验结果（与SPSS软件相同）：　P值　双侧：0.0037192　单侧：0.0018596

图 9-4　甲乙两种血清学方法检查鼻咽癌的检出情况

解：由题意，可将数据整理成表 9-3 与图 9-4 中的四格表形式。由此数据可见，甲法的阳性检出率为$(a+b)/(a+b+c+d)\times100\%=(45+22)/(45+22+6+20)=67/93\times100\%=72.04\%$；同理，乙法的阳性检出率为 51/93×100%=54.84%。检验的目的是看甲法检出率 72.04%对应的总体率 π_1 是否与乙法的检出率 54.84%对应的总体率 π_2 有差异。

H_0：两个总体率相同 $\pi_1=\pi_2$

H_1：两个总体率不同 $\pi_1\neq\pi_2$ α=0.05

对于这种配对四格表，如果两个总体检出率有差异，只可能是甲法阳性而乙法阴性（即 b）或者甲法阴性而乙法阳性（即 c）。所以四个格子中，只需检验 b、c 这两个格子，自然这两个格子的理论频数 $T=(b+c)/2$，根据卡方检验基本公式有：

$$\chi^2=\sum\frac{(A-T)^2}{T}=\frac{\left(b-\frac{b+c}{2}\right)^2}{\frac{b+c}{2}}+\frac{\left(c-\frac{b+c}{2}\right)^2}{\frac{b+c}{2}}=\frac{(b-c)^2}{b+c}\text{，}\ \nu=1$$

当理论频数较小，一般认为 $b+c<40$ 时，需要采用校正卡方检验公式，即

$$\chi^2=\sum\frac{(|A-T|-0.5)^2}{T}=\frac{\left(\left|b-\frac{b+c}{2}\right|-\frac{1}{2}\right)^2}{\frac{b+c}{2}}+\frac{\left(\left|c-\frac{b+c}{2}\right|-\frac{1}{2}\right)^2}{\frac{b+c}{2}}=\frac{(|b-c|-1)^2}{b+c}\text{，}\ \nu=1$$

图 9-5 所显示的工作表共分为 5 个部分：

①第一行为整个检验的标题；

②实际频数；

③卡方检验结果；

④检验临界值；

⑤二项分布精确检验结果。

以上四格表配对资料的卡方检验可按如下步骤进行。

① 按图 9-5 在单元格区域 C6:D7 中键入四格表已知数据，并求周边合计。

② 在单元格 F12、G12、H12 中分别键入自由度“1”、检验水准“0.05”、临界值计算公式“=CHIINV(G12,F12)”。

③ 在单元格 C12 中键入公式“=(D6-C7)^2/(C7+D6)”，计算一般非校正的卡方值。在单元格 D12 中键入“=CHIDIST(C12,$F$12)”，求卡方值对应的 P 值。

④ 在单元格 C13 中键入公式“=(ABS(D6-C7)-1)^2/(C7+D6)”，计算一般校正的卡方值。在单元格 D13 中键入“=CHIDIST(C13,$F$12)”，求校正卡方值对应的 P 值。

⑤ 在单元格 G16 中键入公式“=BINOMDIST(MIN(D6,C7),D6+C7,0.5,1)”，计算二项分布精确检验单侧 P 值。即假定总体概率 $\pi=0.5$，在（b+c）次伯努利试验中，至少出现 b 或 c 次（取较小者）的概率为单侧 P 值。以 2 倍的单侧 P 值作为双侧 P 值。

结果分析：本例 $b+c=28<40$（注意在程序中有红色字提示），故选用连续性校正卡方检验结果，获得 $\chi^2=8.036$，$P=0.005$，按检验水准 $\alpha=0.05$ 拒绝 H_0，认为甲乙两种血清学方法检查鼻咽癌的检出率差别有统计学意义。二项分布精确检验概率双侧 P=0.004，结论与

以上相同。

温馨提示：读者先打开“09-01 卡方检验.xlsm”文件，然后选取“2. 2×2 表配对资料”工作表。在单元格 C6:D7 中，使用滚动条选取或直接键入四格表已知数据，便可获得所需结果。

9.1.3　四格表资料的 Fisher 精确概率检验

在 9.1.1 节中提到，如果总例数 $n \geqslant 40$ 且每个格子理论频数 $T \geqslant 5$，采用 Pearson 卡方检验；如果 $n \geqslant 40$，但理论频数 $1 < T < 5$，则需要采用连续性校正；如果 $T<1$ 或 $n<40$ 时，则需采用四格表资料的 Fisher 精确概率法。此外，用 Pearson 卡方检验或连续性校正计算得到的概率接近检验水准时，也可采用四格表资料的 Fisher 精确概率法。本方法并不属于 χ^2 检验的内容，但可作为四格表 χ^2 检验应用的补充。

该方法的基本思想是：在四格表周边合计不变的条件下，用公式：

$$P=\frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!(a+b+c+d)!}$$

可直接计算出四格表内的 a、b、c、d 四个数据在各种组合下的概率。

因为在周边合计不变的条件下，四格表中任何一个格子改变，其他格子将自动改变（这也就是为什么四格表资料自由度为 1 的原因），所以只需依次增减四格表中任何一个格子（下面的 Excel 程序设定为第 1 行第 1 列格子 a）的数据，便可得到周边合计不变条件下的各种组合的四格表。将小于等于原四格表概率的所有四格表对应的概率相加，所得的和即为双侧概率。包含原四格表概率在内，原表以左为左侧概率，以右为右侧概率。单侧概率一般为左、右侧概率较小者。

例 9-4：为了比较两种药物的驱虫疗效，研究者对 20 名患者进行治疗，甲药 8 人治愈 7 人，乙药 12 人治愈 9 人，其结果见表 9-4。问两种药物的驱虫疗效有无差异？

表 9-4　两种药物的驱虫疗效对比结果

药　物	治愈人数	未治愈人数	总 人 数
甲药	7	1	8
乙药	9	3	12
合计	16	4	20

解：依据题目给定的条件，样本总例数 n=20<40，因此需用 Fisher 精确概率法计算。

H_0：甲乙两药的总体治愈率相等，即 $\pi_1=\pi_2$

H_1：$\pi_1 \neq \pi_2$ $\alpha=0.05$

光盘文件“09-02 Fisher Exact Tests1.xlsx”与“09-03 Fisher Exact Tests2.xlsm”以不同显示方式，进行了 Fisher 精确概率法计算。不过这两个程序均要求四格表第 1 行第 1 列格子 *a* 所对应的行合计或列合计应含有 4 个周边合计中的最小值，否则程序给出提示信息，见图 9-5。

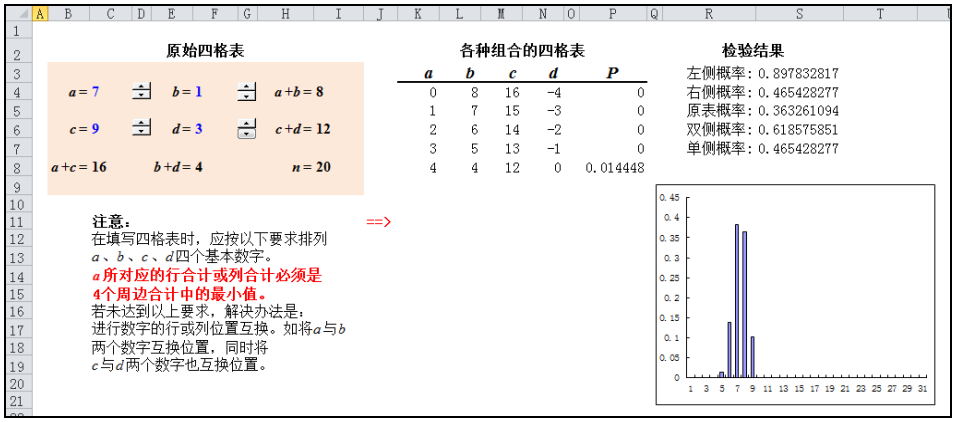


图 9-5　违背四格表排列方式时的提示信息

例 9-4 的表 9-4 中 *a*=7 对应的行合计为 8，列合计为 16，4 个周边合计中的最小值为 4，所以如果按表 9-4 中的 *a*=7、*b*=1、*c*=9、*d*=3 排列方式将数据键入到程序，便会得到图 9-6 的左下方提示信息。这种情况下需要“将 *a* 与 *b* 两个数字互换位置，同时将 *c* 与 *d* 两个数字也互换位置”。对于例 9-4，应修改四格表为 *a*=1、*b*=7、*c*=3、*d*=9。

用 *a*=1、*b*=7、*c*=3、*d*=9 更换图 9-5 中的相应数据（见图 9-6 左上角），获得如图 9-6 所示的结果。

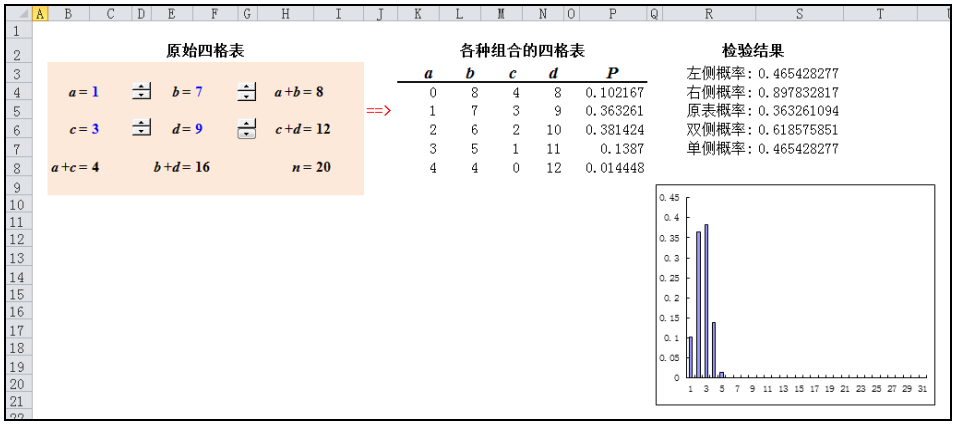


图 9-6　所有可能的四格表及其检验结果

结果分析：图 9-6 中由程序产生了 5 个可能的四格表，其中第 2 个为原四格表（简称原表）。每个四格表的概率显示在第 P 列，5 个四格表概率之和为 1，即

$$0.102167+0.363261+0.381424+0.138700+0.014448=1$$

四格表的概率根据超几何分布概率计算公式计算而得，如原表（程序中 a、b、c、d 四个数字前面有红色“==>”指示）的概率：

$$P_2=\frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!(a+b+c+d)!}=\frac{8!12!4!16!}{1!7!3!9!20!}=0.363261$$

左侧概率为原表以左的所有四格表概率之和，本例左侧概率：

$$P_{左}=0.363261+0.102167=0.465428$$

右侧概率：

$$P_{右}=0.363261+0.381424+0.138700+0.014448=0.897833$$

双侧概率为小于等于原四格表概率的所有四格表对应概率之和，即

$$P_{双侧}=0.102167+0.363261+0.138700+0.014448=0.618576$$

单侧概率一般为左、右侧概率较小者，所以 $P_{单侧}=0.465428$。

本例双侧 P 值为 0.619>0.05，按检验水准 $\alpha=0.05$ 不拒绝 H_0，即尚不能认为甲、乙两药的驱虫疗效有差异。

检验结果的下方是利用所有可能的四格表概率绘制的直条图，可以看出这些概率的分布情况。

温馨提示：该程序使用非常简单，只需键入四格表 4 个数据便可完成计算。该程序将 a、b、c、d 四个值的最大值设定为 20，如果有其中一个值大于 20，则该程序无法运行。

光盘文件“09-03 Fisher Exact Tests2.xlsm”也可以完成四格表精确概率的计算，不过操作略微复杂，因为使用了宏语言，所以打开程序时需“启用宏”。在左上角修改四格表数据，四格表第 1 行第 1 列格子 a 所对应的行合计或列合计应含有 4 个周边合计中的最小值，否则程序给出提示信息。一般情况下，先单击“清除”按钮，再键入新数据，并单击“产生”按钮，来获得所需结果。该程序的特点是将 a、b、c、d 四个数据按四格表方式呈现出来，显示更加直观，如图 9-7 所示。

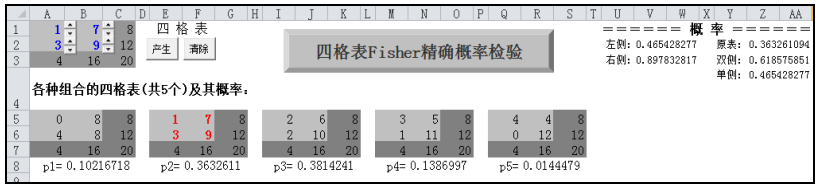


图 9-7　所有可能的四格表

9.1.4 行×列表资料的 χ^2 检验

行×列表（或 $R\times C$ 列联表）的 χ^2 检验主要用于多个率或构成比的比较。可以采用前面介绍的卡方检验基本公式进行计算。

H_0：多个总体率（或构成比）相同

H_1：多个总体率（或构成比）不全相同



注意：行×列表资料的 χ^2 检验，要求理论频数均大于 1，且小于 5 的格子不超过 1/5。否则，需采用其他方法进行检验（如行列表 Fisher 精确概率检验，可采用 SAS 等软件完成），或者采用增大样本含量、合并相邻行或列、删除样本含量少的行或列等办法。



例 9-5：某省观察三个地区的花生污染黄曲霉素 B_1 的情况，见表 9-5。问三个地区花生的黄曲霉素 B_1 污染率有无差别？

表 9-5 某省三个地区花生的黄曲霉素B_1污染率比较

地　区	污染数	未污染数	合　计	污染率（%）
甲	6	23	29	79.3
乙	30	14	44	31.8
丙	8	3	11	27.3
合计	44	40	84	47.6

解：表 9-5 为 3×2 表资料，所以应打开光盘文件“09-01 卡方检验.xlsm”，(启用宏后)选择“6. 3×2 表资料”工作表，将表 9-5 中数据键入到相应格子中，便可获得如图 9-8 所示的计算结果。结果共分为 5 个部分：

①第一行为整个检验的标题；

②实际频数；

③理论频数；

④卡方值计算表；

⑤检验结果。

编制 3×2 表资料卡方检验可按如下步骤进行。

① 在单元格区域 A3:D8 中按图 9-8 键入文本，在单元格区域 B5:C7 键入已知的 3×2 表数据 6、23、…、3。在单元格 D5 中键入公式“=B5+C5”，并将此公式向下拖放填充至单元格 D7，求行合计；在单元格 B8 中键入公式“=B5+B6+B7”，并将此公式向右拖放填

充至单元格 C8 和 D8，分别求列合计和总例数。

返回　3×2表资料的卡方检验

实际频数

	列1	列2	合计
行1	6	23	29
行2	30	14	44
行3	8	3	11
合计	44	40	84

理论频数

	列1	列2	合计
行1	15.19	13.80952381	29
行2	23.05	20.95238095	44
行3	5.762	5.238095238	11
合计	44	40	84

卡方值计算表

	列1	列2	合计
行1	5.56	6.116420361	11.676803
行2	2.097	2.306926407	4.4041322
行3	0.869	0.956277056	1.8256198
合计	8.527	9.379623824	17.906555

检验结果

卡方值	P值	检验水准	自由度	临界值	列联系数
17.90655	0.000129	0.05	2	5.99	0.419184

图 9-8　行×列表的 χ^2 检验

键入已知数据和公式后，选取单元格 A3:D8，单击“开始→自动套用格式（由“文件→选项→自定义功能区”创建），选择一种格式，单击“确定”，设置表格格式。

② 选取单元格区域 A3:D8，并复制，单击单元格 H3，单击“粘贴”按钮。修改“实际频数”为“理论频数”；在单元格 F5 中键入公式“=B$8*$D5/$D$8”，求 3×2 表第一个格子的理论频数，将该公式拖放填充至 3×2 表的其他 5 个格子，求其他格子的理论频数。

③ 选取单元格区域 A3:D8，并复制，单击单元格 H11，单击“粘贴”按钮。修改“实际频数”为“卡方值计算表”。利用卡方检验的基本公式（见 9.1.1 节），在单元格 I13 中键入公式“=(B5-I5)^2/I5”，计算第一个格子的卡方分量值，将此公式向右下拖放填充至单元格 J15，分别计算其他格子的卡方分量值。此时，行合计或列合计之和就是检验所需的卡方值。

④ 按图 9-8 键入单元格区域 A14:F15 中的文本。在单元格 A16 中键入公式“=K16”获取卡方值。在 B16 中键入公式“=CHIDIST(A16,D16)”计算卡方值对应的 P 值。在 C16 中键入检验水准为“0.05”，在 D16 中键入自由度为“2”，在 E16 中键入公式“=CHIINV(C16,D16)”求临界值。

⑤ 根据数学公式 $\sqrt{\text{卡方值}/(\text{卡方值}+\text{总例数})}$，在单元格 F16 中键入公式“=SQRT(K16/(K16+K8))”计算列联系数。当 $R\times C$ 表为双向有序分类变量时，列联系数（contingency coefficient）可定量描述两个分类属性的关联强度，其取值在 0～1 之间，0 表示完全独立，1 表示完全关联。

结果分析：本例 χ^2=17.907，对应 P=0.0001，按 α=0.05 拒绝 H_0，即三个地区花生的黄曲霉素 B_1 污染率差别有统计学意义。本例列联系数=0.4192，为中度相关（一般认为列联系数<0.4 为弱相关，0.4～0.7 为中度相关，≥0.7 为强相关）。

温馨提示：该程序使用非常简单，只需在 3×2 表资料工作表中，键入数据到 B5、C5、B6、C6、B7、C7 六个单元格中，卡方值、对应 P 值、临界值、列联系数即可自动计算获得。

在光盘文件“09-01 卡方检验.xlsm”中还编制了 2×3 表、2×4 表、2×5 表、3×3 表、3×4 表、3×5 表、4×2 表、4×3 表、4×4 表、4×5 表、5×2 表、5×3 表、5×4 表、5×5 表资料的卡方检验工作表。用户只需选用对应的行×列表，并将已知数据键入到实际频数相应单元格中，即可简单执行各种形式行×列表资料的卡方检验。

9.2 频数分布的拟合优度检验

卡方检验不仅可用来检验名义分类数据的两组或多组样本率（或构成比）之间差异有无统计学意义，还可用来检验区间数据多个样本对应的总体方差是否齐同（见 8.5 节），此外，也可用来检验频数分布的拟合优度。

频数分布的拟合优度（goodness of fit）检验就是判定某实际频数分布是否符合某一理论分布。其基本思想是按某理论分布计算出理论频数，利用χ^2 值比较实际频数和理论频数的吻合程度，从而推断频数分布的拟合优度。

H_0：资料对应总体符合某分布

H_1：资料对应总体不符合某分布

$$\chi^2=\Sigma\frac{(A-T)^2}{T}\text{，}\ v\text{=组数–计算理论频数 }T\text{ 时所用参数的个数}$$

式中，A 是实际频数，T 是理论频数，组数是参加比较的实际频数的组数。该统计量服从自由度为 v 的χ^2 分布。注意，和前面 Pearsonχ^2 检验一样，要求每个格子的理论频数不小于 5。如果小于 5，应考虑相邻实际频数合并。

χ^2 值实质上是以理论频数为基数的相对误差，反映了实际频数与理论频数差别的程度。如果该实际分布服从某理论分布，则差别不应很大。

科学研究中常用的分布包括正态分布、二项分布、Poisson 分布、均匀分布等，下面分别给出其 Excel 实现拟合优度检验的方法。

9.2.1 正态分布

正态分布是科学研究数据中最常见、最重要的一种分布，它是许多统计推断的基础，也是许多分布的极限分布。其累积分布函数（cumulative distribution function，CDF）公式为：

$$F(X)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{X}\exp\left(-\frac{(X-\mu)^2}{2\sigma^2}\right)\mathrm{d}X=\int_{-\infty}^{X}f(X)\mathrm{d}X$$

式中，X 为该变量的实际取值，$F(X)$为 X 值对应的累积分布概率，π=3.14159 为圆周率，$\exp(\bullet)$即$e^{(\bullet)}$，e 为自然对数的底，μ和σ为正态分布的两个参数，即分别为总体均数（μ，即位置参数）和总体标准差（σ，即形状参数），$f(X)$ 为正态分布的概率密度函数（probability density function, PDF）。累积分布函数与概率密度函数的图示见图 9-9，其中 X 值以左正态分布曲线下的面积值就是 $F(X)$。当 $X=\infty$时，$F(X)=1$，也就是正态分布曲线 $f(X)$ 下的全部面积为 1。

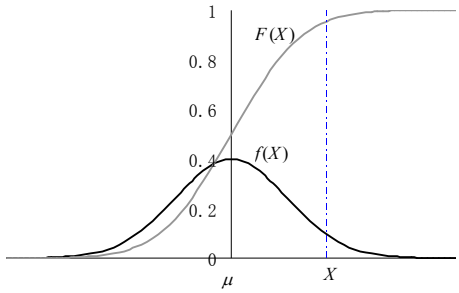


图 9-9　正态分布概率函数图示

$F(X)$的 Excel 计算公式的格式为“NORMDIST（X，均数，标准差，1）”，其中“1”含义为“true”，表示需要累积；如果选“0”或“false”，表示不需要累积，此时计算的就是 $f(X)$，即正态分布曲线的轨迹。

正态分布拟合优度检验计算理论频数 T 时所用参数个数为 2，即均数 μ 和标准差 σ。

例 9-6：某研究者测得 120 例 20～50 岁正常人血浆结合 ^{125}碘-三碘甲状腺原氨酸（^{125}I-T_3）树脂摄取比值，其频数分布见图 9-10 中的第 A 列至第 D 列（也见光盘文件“09-04 正态分布拟合优度检验.xlsx”）。问该资料是否服从正态分布？

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	正态分布拟合优度检验														
2	组段下限	组段上限	组中值X	实际频数f	fX	fX^2	正态分布概率P_X	理论频数T	卡方分量$(f-T)^2/T$						
3	0.75	0.80	0.775	1	0.775	0.601	0.009297	1.1			均数	1.0054167		$\sum fX$	120.650
4	0.80	0.85	0.825	2	1.650	1.361	0.028184	3.4	0.413		标准差	0.087278		$\sum fX^2$	122.210
5	0.85	0.90	0.875	13	11.375	9.953	0.076077	9.1			例数	120		$\sum P_X$	1.000
6	0.90	0.95	0.925	15	13.875	12.834	0.149175	17.9	0.470		组距	0.05		$\sum T$	120.000
7	0.95	1.00	0.975	26	25.350	24.716	0.212524	25.5	0.010						
8	1.00	1.05	1.025	26	26.650	27.316	0.220004	26.4	0.006		卡方统计量	3.2456			
9	1.05	1.10	1.075	18	19.350	20.801	0.165490	19.9	0.174		自由度	5			
10	1.10	1.15	1.125	15	16.875	18.984	0.090447	10.9	1.584		检验水准α	0.05			
11	1.15	1.20	1.175	3	3.525	4.142	0.035910	4.3	0.588		P值	0.6622			
12	1.20	1.25	1.225	1	1.225	1.501	0.012892	1.5			临界值	11.0705			

图 9-10　正态分布拟合优度检验

解：正态分布拟合优度检验可按下列步骤进行。

① 按图 9-10 建立工作表。在单元格区域 A1:I2、K3:K12、N3:N6 中键入文字、符号。

在单元格 A3 中键入第一个组段的下限值 0.75，在 L6 中键入组距 0.05。

② 在单元格 A3 中键入公式“=A3+$L$6”，向下拖放至 A12；在单元格 B3 中键入公式“A3+$L$6”，计算上限值。

③ 在单元格 C3 中键入公式“AVERAGE(A3:B3)”，向下拖放至 C12，计算组中值，在单元格区域 D3:D12 中键入已知的各组频数。

④ 在单元格 E3 和 F3 中分别键入公式“=D3*C3”和“=D3*C3^2”，选取单元格区域 E3:F3，将此公式向下拖放填充至第 12 行，完成 fx 和 fx^2 的计算。

⑤ 在单元格 L5 中键入公式“=SUM(D3:D12)”求总例数，在单元格 O3、O4 中分别键入公式“=SUM(E:E)”、“=SUM(F:F)”，计算 $\sum fx$ 和 $\sum fx^2$。

⑥ 在单元格 L3 中键入公式“=O3/L5”，计算频数资料的均数；在单元格 L4 中键入公式“=SQRT((O4-O3^2/L5)/(L5-1))”，计算频数资料的标准差。数学计算公式如下：

$$\bar{X}=\frac{\sum f_i x_i}{\sum f_i},\quad S=\sqrt{\frac{\sum f_i x_i^2-\frac{(\sum f_i x_i)^2}{\sum f_i}}{\sum f_i-1}}$$

⑦ 计算各组段的正态分布曲线下概率。因为正态分布是连续分布，故某一区间内曲线下面积可以用累积分布之差来计算。在单元格 G3 中键入公式“=NORMDIST(B3,$L$3,$L$4,1)”，计算树脂摄取比值小于 0.8 的概率；在单元格 G4 中键入公式“NORMDIST(B4,$L$3,$L$4,1)-NORMDIST(B3,$L$3,$L$4,1)”，计算树脂摄取比值在 0.80～0.85 之间的概率。将单元格 G4 中的公式向下拖放填充至 G11。在单元格 G12 中键入公式“1-SUM($G$3:G11))”，计算树脂摄取比值大于等于 1.2 的概率。

⑧ 在单元格 H3 中键入公式“G3*$L$5”，计算正态分布的理论频数 T。将此公式向下拖放填充至单元格 H12，完成 T 的计算。在 O5、O6 中分别键入公式“=SUM(G:G)”、“=SUM(H:H)”，计算 $\sum P_X=1$、$\sum T=120$，主要起验证作用，这两个值并不参与后面计算。

⑨ 理论频数小于 5 的行需与相邻行合并，使理论频数之和大于 5。本例在前两个组段（见单元格 H3、H4）理论频数小于 5，合并后仍小于 5，因此将单元格 I3、I4、I5 合并。在合并的单元格 I3 中键入公式“=(SUM(D3:D5)-SUM(H3:H5))^2/SUM(H3:H5)”，计算三个组段合并的卡方分量。同理，合并单元格 I11、I12，在合并的单元格 I11 中键入公式“=(SUM(D11:D12)-SUM(H11:H12))^2/SUM(H11:H12)”，计算两个组段合并的卡方分量。在单元格 I6 中键入公式“=(H6-D6)^2/H6”，计算第 4 个组段的卡方分量，将此公式向下拖放填充至单元格 I10，完成其余卡方分量的计算。

⑩ 在单元格 L8 中键入公式“=SUM(I:I)”，获得卡方值；在单元格 L9 中键入公式“=COUNT(I:I)-2”，计算自由度。本例合并后实际有 7 组数参加计算，用到正态分布理论频数计算有两个参数，故自由度 7-2=5。在单元格 L10 中指定检验水准 α，本例指定“0.05”；

在单元格 L11 中键入公式“=CHIDIST(L8,L9)”，求卡方值对应的 P 值；在单元格 L12 中键入公式“=CHIINV(L10,L9)”，计算χ^2临界值。

结果分析：本例χ^2=3.245，P=0.662，按 α=0.05 不拒绝 H_0，可认为该资料服从正态分布。

温馨提示：光盘文件“09-04 正态分布拟合优度检验.xlsx”在以上叙述的基础上，又作了部分修订，目的是让程序使用更加简洁、通用。修改要点如下：

①在单元格 A3 中修改第一组频数的下限值，在单元格 L6 中修改组距，将单元格 A4 公式向下拖放至所需位置；

②修改第 D 列实际频数值；

③根据第 H 列的理论频数，将理论频数小于 5 的第 H 列单元格进行“合并”，并仿照上面第 ⑨ 步进行修改；

④如果必要，可对 L10 的检验水准进行修订。程序中黑色字体部分的公式会自动根据修改数据变化而变化。

9.2.2　二项分布

二项分布的概率质量函数（probability mass function，PMF）为：$P(X=x)=\binom{n}{x}\pi^x(1-\pi)^{n-x}$，$X$ =0, 1, …, x, …, n，其中 n 为试验次数，π 为总体率，X 为随机变量，取值为 x。二项分布的 pmf 图形如图 9-11 所示。

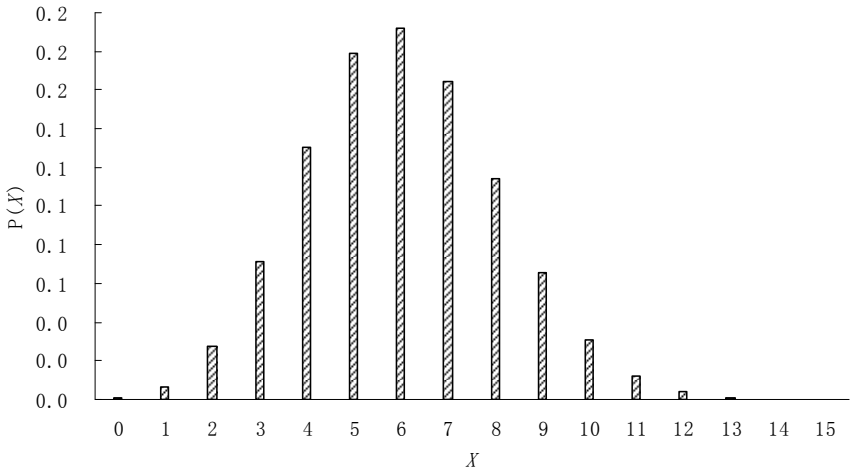


图 9-11　二项分布

二项分布拟合优度检验计算理论频数 T 时所用参数个数为 2（即试验次数 n 和总体率 π）。

例 9-7：某地爆发流行性细菌性痢疾，调查了 288 户四口之家，其每户发病人数的频数分布见图 9-12（也见光盘文件“09-05 二项分布拟合优度检验.xlsx”）的单元格区域 A3:B7。问该资料是否服从二项分布？

	A	B	C	D	E	F	G	H	I	J	K	L
1	二项分布拟合优度检验											
2	每户发病人数 X	观察家庭数 A	发病人数	二项分布概率 $P(X=x)$	理论频数 T	卡方分量 $(A-T)^2/T$						
3	0	167	0	0.439543	126.6	12.90		卡方统计量	81.00		发病率 π	0.185764
4	1	51	51	0.401118	115.5	36.04		自由度	2		$\sum A$	288
5	2	50	100	0.137270	39.5	2.77		检验水准 α	0.05		共计发病人数	214
6	3	17	51	0.020878	6.0	29.29		P 值	3E-18		$\sum P$	1.000000
7	4	3	12	0.001191	0.3			临界值	5.9915		$\sum T$	288.0

图 9-12　二项分布拟合优度检验

解：二项分布拟合优度检验可按下列步骤进行。

① 按图 9-12 建立工作表。在单元格区域 A1:F2、H3:H7、K3:K7 中键入文字、符号，并在单元格区域 A3:A7 中键入可能的几种每户发病人数 X，在 B3:B7 中键入与 A3:A7 对应的观察家庭数 A。单元格 L4 为总的观察家庭数，可键入公式“=SUM(B:B)”获得。

② 在单元格 C3 中键入公式“=A3*B3”，计算各组发病人数，将此公式向下拖放填充至单元格 C7。在单元格 L5 中键入公式“=SUM(C:C)”，求总的发病人数。

③ 在单元格 L3 中键入公式“=L5/(L4*4)”，计算二项分布的参数发病率 π。

④ 根据函数“BINOMDIST(例数 x，试验次数 n，发病率 π，0)”计算二项分布概率。在单元格 D3 中键入公式“=BINOMDIST(A3,$A$7,$L$3,0)”，将此公式向下拖放填充至单元格 D7。

⑤ 根据“二项分布概率×观察家庭总数”，在单元格 E3 中键入公式“=D3*$L$4”，计算理论频数，将此公式向下拖放填充至单元格 E7。

⑥ 根据卡方分量等于 $(A-T)^2/T$，在单元格 F3 中键入公式“=(B3-E3)^2/E3”，计算相应的卡方分量，将此公式向下拖放填充至单元格 F5。因为在 X=4 时的理论频数为 $0.3<5$，故将后两行合并。在合并后的单元格 F6 中修改计算公式“=(B6+B7-E6-E7)^2/(E6+E7)”。

⑦ 在单元格 I3 中键入公式“=SUM(F:F)”，获得卡方值；在单元格 I4 中键入公式“=COUNT(F:F)-2”，计算自由度。本例有 5 组参加计算，合并后实际只有 4 组，用了两个参数，故自由度=4–2=2。在单元格 I5 中指定检验水准 α，本例指定“0.05”；在单元格 I6 中键入公式“=CHIDIST(I3,I4)”，求卡方值对应的 P 值；在单元格 I7 中键入公式“=CHIINV(I5,I4)”，计算 χ^2 临界值。

结果分析：本例获得 χ^2=81.00，$P=3\times10^{-18}$，按 α=0.05 拒绝 H_0，可认为本次调查的四口之家内发病人数的分布不服从二项分布。

温馨提示：修改要点如下：

①更换第 A、B 列数据；

②将第 C、D、E 列公式下拉到适合自己数据的位置；

③根据第 E 列的理论频数，将理论频数小于 5 的单元格进行“合并”；

④如果必要，可对单元格 I5 的检验水准进行修订。

9.2.3　Poisson 分布

Poisson 分布的概率质量函数公式：

$$P(X=x)=\frac{e^{-\lambda}\lambda^{x}}{x!}，\ X=0,\ 1,\ \cdots,\ x,\ \cdots$$

式中，随机变量 X 的取值为整数 x，e＝2.718 为自然对数的底，λ 为均数，是 Poisson 分布的唯一参数。

Poisson 分布拟合优度检验在计算理论频数 T 时，所用参数个数为 2（即均数 λ 和总例数 n）。

例 9-8：300 个单位容积内的细菌计数结果如图 9-13 的第 A、B 两列所示，问此资料是否服从 Poisson 分布？

	A	B	C	D	E	F	G	H	I	J	K	L
1	Poisson分布拟合优度检验											
2	单位容积内细菌数X	观察频数A	观察细菌数$X\cdot A$	Possion分布概率$P(X=x)$	理论频数T	卡方分量$(A-T)^2/T$						
3	0	26	0	0.082910	24.9	0.051						
4	1	51	51	0.206446	61.9	1.930		卡方统计量	4.471		均数μ	2.49
5	2	84	168	0.257025	77.1	0.616		自由度	5		$\sum A$	300
6	3	70	210	0.213331	64.0	0.563		检验水准α	0.05		$\sum XA$	747
7	4	42	168	0.132798	39.8	0.117		P值	0.483729		$\sum P(X=x)$	1.000000
8	5	15	75	0.066134	19.8	1.181		临界值	11.0705		$\sum T$	300.0
9	6	9	54	0.027445	8.2	0.013						
10	7	3	21	0.013911	4.2							

图 9-13　Poisson 分布拟合优度检验

解：Poisson 分布拟合优度检验可按下列步骤进行。

①按图 9-13（也见光盘文件“09-06 Poisson 分布拟合优度检验.xlsx”）建立工作表。在单元格区域 A1:F2、H4:H8、K4:K8 中键入文字、符号，并在单元格区域 A3:A10 中键入可能的单位容积内细菌数 X，在 B3:B10 中键入与 A3:A10 对应的观察频数 A。单元格 L5 为总的单位容积内细菌数 $\sum A$，可键入公式“=SUM(B:B)”获得。

②在单元格 C3 中键入公式“=A3*B3”，计算各组观察细菌数，将此公式向下拖放填充至单元格 C10。在单元格 L6 中键入公式“=SUM(C:C)”，求总的观察细菌数 $\sum XA$。

③在单元格 L4 中键入公式“=L6/L5”，计算 Poisson 分布的均数 μ。

④根据函数格式“POISSON(变量 X=x，参数 λ，0)”计算 Poisson 分布概率。在单

元格 D3 中键入公式“=POISSON(A3,$B$13,0)”，将此公式向下拖放填充至单元格 D9。在单元格 D10 中键入公式“=1–SUM(D3:D9)”，计算单位容积内细菌数大于等于 7 的概率。

⑤根据“Poisson 分布概率×观察总频数”，在单元格 E3 中键入公式“D3*$L$5”，计算理论频数，将此公式向下拖放填充至单元格 E10。

⑥根据卡方分量=$(A-T)^2/T$，在单元格 F3 中键入公式“=(B3–E3)^2/E3”，计算相应的卡方分量，将此公式向下拖放填充至单元格 F10。

⑦观察是否有某行的理论频数小于 5，若有则需要将相近的行进行合并。本例在 X=7 时的理论频数为 4.2＜5，故将后两行合并（合并方法同前）。

⑧在单元格 I4 中键入公式“=SUM(F:F)”，计算卡方值；在单元格 I5 中键入公式“=COUNT(F:F)–2”，计算自由度。本例有 8 组数参加计算，合并后实际只有 7 组，用了 2 个参数，故自由度=7–2=5；在单元格 I6 中指定检验水准α，本例指定“0.05”；在单元格 I7 中键入公式“=CHIDIST(I4,I5)”，求卡方值对应的 P 值；在单元格 I8 中键入公式“=CHIINV(I6,I5)”，计算χ^2临界值。

结果分析：本例χ^2=4.471，P=0.484，按α=0.05 不拒绝 H_0，可认为本资料服从 Poisson 分布。

温馨提示：修改要点如下：

①更换第 A、B 列数据；

②根据第 E 列的理论频数，将理论频数小于 5 的单元格进行“合并”；

③如果必要，可对单元格 I6 的检验水准进行修订。

9.2.4 均匀分布

均匀分布即等概率分布，指变量取值的机会均等。均匀分布概率函数公式为 $P(X)=1/n$，式中 n 为变量取值的个数。

均匀分布拟合优度检验在计算理论频数 T 时所用参数只有 1 个，即 n。

例 9-9：一粒骰子有 6 个面，分别刻有 1、2、3、4、5 和 6，共投掷 60 次，各数字出现的频数如图 9-14（也见光盘文件“09-07 均匀分布拟合优度检验.xlsx”）的第 A、B 两列所示，问这粒骰子投掷的频数是否均匀？

解：均匀分布拟合优度检验可按下列步骤进行。

①按图 9-14 建立工作表。第 A、B 列为原始数据。

②在单元格 C3 中键入公式“=1/COUNT(A:A)”，计算均匀分布的概率。将此公式向下拖放填充至 C8。

③在单元格 D3 中键入公式“=C3*$K$3”，计算均匀分布的理论频数。将此公式向下

拖放填充至 D8。

	A	B	C	D	E	F	G	H	I	J	K
1	均匀分布拟合优度检验										
2	每面取值 X	观察频数 A	均匀分布概率 P_X	理论频数 T	卡方分量 $(A-T)^2/T$						
3	1	8	0.166666667	10	0.4		卡方统计量	7.400		$\sum A$	60
4	2	8	0.166666667	10	0.4		自由度	5		$\sum P_X$	1
5	3	5	0.166666667	10	2.5		检验水准α	0.05		$\sum T$	60
6	4	10	0.166666667	10	0.0		P值	0.193			
7	5	14	0.166666667	10	1.6		临界值	11.070			
8	6	15	0.166666667	10	2.5						

图 9-14　均匀分布拟合优度检验

④在单元格 E3 中键入公式“=(B3–D3)^2/D3”，计算相应的卡方分量$(A–T)^2/T$。将此公式向下拖放填充至 E8。

⑤在单元格 K3 中键入公式“=SUM(B:B)”，计算观察频数的总频数$\sum A$。在单元格 K4 中键入公式“=SUM(C:C)”，计算总概率$\sum P_X$；在单元格 K5 中键入公式“=SUM(D:D)”，计算理论频数合计$\sum T$。

⑥在单元格 H3 中键入公式“=SUM(E:E)”计算卡方值，在单元格 H4 中键入公式“=COUNT(A:A)–1”计算自由度，本例有 6 组数据参加计算，计算理论频数仅用到总例数 1 个参数，故自由度=6–1=5。在单元格 H5 中指定检验水准α，本例指定“0.05”。在单元格 H6 中键入公式“=CHIDIST(H3,H4)”，求卡方值对应的 P 值。在单元格 H7 中键入公式“=CHIINV(H5,H4)”，计算χ^2临界值。

结果分析：本例χ^2=7.4，P=0.193，按 α=0.05 水准不拒绝 H_0，可认为这粒骰子投掷的频数分布呈均匀分布。

温馨提示：修改要点如下：

①更换第 A、B 列数据；

②如果必要，可对单元格 I6 的检验水准进行修订。

9.3　非参数检验

第 8 章介绍的 t 检验与方差分析，都是在已知总体分布形状（如正态分布）的条件下，对未知总体参数进行统计推断的方法，这类方法称为参数法。在实际工作中，结果变量尽管是区间数据，但数据达不到参数法的要求（如服从正态分布、各组方差齐同），则需要采用非参数检验方法。

上面 9.1 节介绍的卡方检验，结果变量是名义数据。如果要考虑结果变量的顺序关系，也需要采用非参数检验。

非参数检验实际上是将结果变量从小到大排列顺序，并从 1 到 n（观察总例数或对子数）按自然数编顺序号（秩次，rank），对于相同数值或等级（tie）取平均秩次（mean rank）。例如，11 只老鼠存活天数分别为：4、10、7、50、3、15、2、9、13、>60、>60，则秩次为 3、6、4、9、2、8、1、5、7、10.5、10.5。第 7 只老鼠仅存活 2 天，存活时间最少，因此秩次为 1；其次少的是第 5 只老鼠，仅活了 3 天，因此秩次为 2；第 10 与第 11 只老鼠均存活了 60 天以上，所以秩次应该是 10、11，但按照相同秩次取平均秩次的原则，所以各取平均 10.5。又如 7 名肺炎病人的治疗结果为：治愈、治愈、死亡、无效、治愈、有效、治愈，假定危险程度由弱到强依次是：治愈、有效、无效、死亡，则 7 名肺炎病人相应秩次为 2.5、2.5、7、6、2.5、5、2.5。共有 4 个“治愈”，秩次应该为 1、2、3、4，所以它们的平均秩次为 2.5。

下面将要介绍的基于秩次的方法有：Wilcoxon 符号秩检验、两样本 Mann-Whitney U 检验、多样本 Kruskal-Wallis 秩和检验、随机区组资料 Friedman 秩和检验。

Excel 没有直接提供上述非参数检验的数据分析工具，需要利用函数和公式实现。

9.3.1 配对资料 Wilcoxon 符号秩检验

配对资料 Wilcoxon 符号秩检验（Wilcoxon signed ranks test）的基本思想是：先求出每对观测值的差值 d，按$|d|$的大小编秩，分别求出正、负秩次之和。如果两观测结果相同，差值的“正秩和”与“负秩和”应相差不大，差值的总体中位数为 0。反之，如果差别太大，就拒绝分布位置相同的假设。

H_0：差值的总体中位数为 0

H_1：差值的总体中位数不为 0



注意：编秩时依照差值的绝对值从小到大编秩，再将差值的正负号标在秩次之前。若有绝对值相等者（相同，tie），取平均秩次；零差值不参与编秩，相应的对子数应减少。

取正秩和与负秩和较小者作为检验统计量 T，小样本时查配对比较的秩和检验界值表，该表是根据差值秩和的分布，直接计算出来的。界值的判断标准为：若下限<T<上限，P 值>表中概率值，若 T≤下限或 T≥上限，则 P 值≤表中概率值。样本较大（如对子数 n≥25）时，差值的秩和分布逐渐逼近

$$\text{均数 } \mu=\frac{n(n+1)}{4},\ \text{方差 } \sigma^2=\frac{n(n+1)(2n+1)}{24}$$

的正态分布，故可用正态近似法进行检验。即

$$z=\frac{|T-\mu|}{\sigma}=\frac{|T-\mu|}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$$

$$z_c=\frac{|T-n(n+1)/4|}{\sqrt{\frac{n(n+1)(2n+1)}{24}-\frac{\sum(t_j^3-t_j)}{48}}}$$

其中 n 为对子数，T 为检验统计量，即正秩和与负秩和中的较小者。

例 9-10：对 10 名健康人分别用离子交换法与蒸馏法，测得尿汞值如表 9-6 所示。问两种方法的测定结果有无差别？

表 9-6　10 名健康人用离子交换法与蒸馏法测定尿汞值（μg/l）

样 品 号	1	2	3	4	5	6	7	8	9	10
离子交换法	0.5	2.2	0.0	2.3	6.2	1.0	1.8	4.4	2.7	1.3
蒸 馏 法	0.0	1.1	0.0	1.3	3.4	4.6	1.1	4.6	3.4	2.1

	A	B	D	E	F	H	I	J
1	配对资料Wilcoxon符号秩检验							
2						α	0.05	
3	原始数据		差值及秩次			H0:	μ1 = μ2	
4	离子交换法	蒸馏法	差值d	\|d\|	秩次	检验结果		
5	0.5	0	-0.5	0.5	-2	对子数	10	
6	2.2	1.1	-1.1	1.1	-7	n	9	
7	0	0	0	0		负秩和T⁻	26.5	平均负秩 5.3
8	2.3	1.3	-1	1	-6	正秩和T⁺	18.5	平均正秩 4.625
9	6.2	3.4	-2.8	2.8	-8	T=	18.5	
10	1	4.6	3.6	3.6	9		下限	上限
11	1.8	1.1	-0.7	0.7	-3.5	T临界值	5	40
12	4.4	4.6	0.2	0.2	1	结论：	P值大于 0.05	
13	2.7	3.4	0.7	0.7	3.5		不拒绝H0	
14	1.3	2.1	0.8	0.8	5			
15						n≥25时的正态近似		
16						均数μ	22.5	
17						标准差σ	8.4410	
18								
19						z值	0.4739	
20						P值(双侧)	0.6356	
21						P值(单侧)	0.3178	

图 9-15　配对资料 Wilcoxon 符号秩检验

解：配对资料 Wilcoxon 符号秩检验可按下列步骤进行。

① 按图 9-15（也见光盘文件“09-08 配对资料 Wilcoxon 符号秩检验.xlsx”）将表 9-6 中数据输入到工作表的第 A、B 列中。在 I2 单元格中，选取检验水准 α。单击单元格 I2 后，会在右侧出现下拉箭头（根据数据“有效性”制成），单击下拉箭头可从中选取 0.1、0.05、0.02、0.01。

② 计算每对数据的差值和其绝对值|d|，在单元格 D5 中输入公式“=IF(AND(A5<>"",

B5<>""),B5-A5,"")”，在单元格 E5 中输入公式“=IF(AND(A5<>"",B5<>""),ROUND(ABS(D5),6),"")”。选取单元格区域 D5:E5，向下拖放填充至第 54 行，完成差值及其绝对值的计算。

③按|*d*|从小到大编秩，并将 *d* 的正负号标在秩次之前。*d* 为 0 不参加编秩。采用 IF 和 COUNTIF 函数，在单元格 F5 中键入公式“=IF(AND(A5<>"",B5<>""),IF(D5>0,COUNTIF(E:E,"<"&E5)-COUNTIF(E:E, "=0")+(1+COUNTIF(E:E, "="&E5))/2,IF(D5<0,-(COUNTIF(E:E,"<"&E5)-COUNTIF(E:E, "=0")+(1+COUNTIF(E:E, "="&E5))/2),"")),"")”，对第 1 对数据的差值编秩次。单击单元格 F5，将此公式向下拖放填充至单元格 F54，完成其余差值的编秩。

④在单元格 I5 中键入公式“=COUNT(E:E)”，计算总对子数；在单元格 I6 中键入公式“=COUNT(E:E)-COUNTIF(E:E, "=0")”，计算删除差值为 0 对子后的对子数。

⑤在单元格 I7 中键入公式“=ABS(SUMIF(F:F,"<0"))”，计算负秩和，本例为“–18.5”。在单元格 I8 中键入公式“=SUMIF(F:F,">0")”，计算正秩和 T^+，本例为“26.5”。从正、负秩和中选择较小者作为检验统计量 *T*，公式为“=MIN(I7,I8)”。

⑥在单元格 I11 中键入公式“=INDEX($M$5:$P$54,$I$6,MATCH($I$2,$M$4:$P$4,0))”，在 *T* 临界值区域查找相应下限临界值；在单元格 J11 中键入公式“=INDEX($S$5:$V$54,$I$6,MATCH($I$2,$S$4:$V$4,0))”，在 *T* 临界值区域查找相应上限临界值。

⑦在单元格 I12、I13 和 J12 中，根据统计量 *T* 与查找到的临界值下假设检验结论。

⑧如果 $n \geqslant 25$，则可采用正态近似的检验结果。在单元格 I16 中键入公式“=I6*(I6+1)/4”，计算均数 μ；在单元格 I17 中键入公式“=SQRT(I6*(I6+1)*(2*I6+1)/24)”，计算标准差 σ。在单元格 I19 中键入公式“ABS(I9-I16)/I17”，计算检验统计量 *z* 值；在单元格 I20 中键入公式“= 2*NORMDIST($I$9,$I$16,$I$17,1)”，计算 *z* 值对应的双侧 *P* 值；在单元格 I21 中键入公式“= NORMDIST($I$9,$I$16,$I$17,1)”，计算 *z* 值对应的单侧 *P* 值。为了避免结果选择错误，所以当参与比较的对子数小于 25 时，正态近似 *z* 检验结果将呈浅灰色显示，如本例参与比较数据对子数仅 9 个，所以 *z* 检验相关结果呈浅灰色。

结果分析：结果如图 9-15 所示，本例检验统计量 *T*=18.5，按 *n*=9 查所附的界值表，双侧 0.05 的界值为（5～40），*T*=18.5 均落在此范围内，故 *P*>0.05，不能拒绝 H_0。尚不能认为两种方法的结果有差别。

温馨提示：该程序最多可处理 50 个对子。不仅适用于配对资料的比较，而且也适用于单样本与总体中位数的比较。修改要点如下：

①更换第 A、B 列数据；

②如果必要，可对 I2 的检验水准进行修订。

请注意：本程序的 *z* 值计算公式没有考虑相同秩次问题，因此，如果相同秩次较多时，获得的结果会与常用统计软件有所不同，建议使用常用统计软件如 SAS、SPSS 等输出的结果。

9.3.2　两独立样本非参数检验

两样本 Wilcoxon 秩和检验与 Mann-Whitney U 检验可用来比较两个独立样本，推断两个总体的分布位置是否有差异。

H_0：两个总体的分布位置相同

H_1：两个总体的分布位置不同

其基本思想是：将两样本混合编秩次，若两组样本的总体分布相同，则两样本的秩和在 $n_1=n_2$ 时应大致相等；当 $n_1\neq n_2$ 时则两样本的秩和应与各样本含量成比例，即平均秩次（mean ranks）相同。反之，当两总体分布位置不同时，平均秩次有差异。



注意：编秩时若有相同值则取平均秩次。记样本含量少的一组获得的秩和为统计量 T，相应样本含量记为 n_1。即规定 $n_1\leqslant n_2$。如果 $n_1=n_2$，可取任一组别的秩和为统计量 T，但为了方便，仍取 n_1 对应的秩和为统计量 T

小样本时查“两样本比较秩和检验的 T 界值表”，该表根据秩和的分布直接计算而得。界值的判断标准同前。即若下限<T<上限，P 值>表中概率值；若 $T\leqslant$下限或 $T\geqslant$上限，则 P 值≤表中概率值。

在大样本时，秩和的分布逐渐逼近

$$\text{均数 } \mu=\frac{n_1(n_1+n_2+1)}{2},\ \text{方差 } \sigma^2=\frac{n_1n_2(n_1+n_2+1)}{12}$$

的正态分布，故可用正态近似法进行检验。

$$z=\frac{|T-\mu|}{\sigma}=\frac{|T-\mu|}{\sqrt{\frac{n_1n_2(n_1+n_2+1)}{12}}}$$

Mann-Whitney U 检验是将某一组的每一个值与另一组的所有值进行比较，如果较大则记为 1，相等记为 0.5，较小记为 0，其比较结果之和就是 U 值。如果令第一组每一个值与第二组的所有值进行比较的结果之和为 U_1，如果令第二组每一个值与第一组的所有值进行比较的结果之和为 U_2，则有：

$$U_1=n_1n_2+\frac{n_2(n_2+1)}{2}-T_2,\quad U_2=n_1n_2+\frac{n_1(n_1+1)}{2}-T_1$$

并且一定有：

$$U_1+U_2=n_1n_2$$

以此可用于核对是否计算正确。

U_1、U_2 中较小者小于下限界值，或 U_1、U_2 中较大者大于上限界值，则 P 值≤表中概

率值，有理由拒绝 H_0，表明差异有统计学意义。否则不拒绝 H_0，表明差异无统计学意义。

例 9-11：某实验室观察局部温热治疗小鼠植性肿瘤的疗效，以生存日数作为观察指标，资料见表 9-7。试检验两组小鼠生存日数有无差别？

表 9-7　两组小鼠发癌后生存日数

实验组生存日数	10	12	15	15	16	17	18	20	23	90		
对照组生存日数	2	3	4	5	6	7	8	9	10	11	12	13

两样本Wilcoxon秩和检验与Mann-Whitney U检验

α 0.05　$H_0: \mu_1 = \mu_2$

原始数据		秩次		Wilcoxon秩和检验结果			Mann-Whitney U检验结果	
第一组	第二组	第一组	第二组	第一组T_1	第二组T_2	统计量T	U_1	U_2
10	2	9.5	1	170	83	170	115	5
12	3	12.5	2	n_1	n_2			
15	4	15.5	3	10	12		U临界值	下限 29
15	5	15.5	4	平均秩次1	平均秩次2		结论：	P值小于等于 0.05
16	6	17	5	17	6.916666667			拒绝H0
17	7	18	6		下限	上限		
18	8	19	7	T临界值	84	146		
20	9	20	8	结论：	P值小于等于 0.05		正态近似结果	
23	10	21	9.5		拒绝H0		均数μ	115
90	11	22	11				标准差σ	15.16575089
	12		12.5				z值	3.626592604
	13		14	结果验证	$n_1 n_2$	U_1+U_2	P值(双侧)	0.000287186
					120	120	P值(单侧)	0.000143593

图 9-16　两独立样本非参数检验

解：在图 9-16（也见光盘文件“09-09 两独立样本非参数检验.xlsx”）中，结果主要分为三个部分，分别是两独立样本 Wilcoxon 秩和检验、Mann-Whitney U 检验和正态近似检验结果。

可按下列步骤实现：

① 按图 9-16 将表 9-7 中数据输入到工作表的第 A、B 列中。注意：为了不引起程序计算误会，最好将样本含量较少的一组放在第 A 列。

② 在单元格 D5 中输入公式“=IF(A5<>"",(COUNTIF($A:$B,"<"&A5)+(1+COUNTIF($A:$B, "="&A5))/2),"")”，对单元格 A5 中的数值在两组数据中排秩次。单击单元格 D5，将此公式向右拖放填充至 E5；再选取单元格区域 D5:E5，向下拖动填充至第 54 行，完成两组数据混合编秩。

③ 在单元格 G5 中键入公式“=SUM(D:D)”，计算第一组秩次之和秩和 T_1；在单元格 H5 中键入公式“=SUM(E:E)”，计算第二组秩次之和秩和 T_2；在单元格 I5 中键入公式“=IF(COUNT(D:D)<COUNT(E:E),G5,H5)”，以较小样本组的秩和为统计量 T。

④ 在单元格 G7 中键入公式“=MIN(COUNT($D:$D),COUNT($E:$E))”，获得例数 n_1；在单元格 H7 中键入公式“=MAX(COUNT($D:$D),COUNT($E:$E))”，获得例数 n_2。在单元

格 G9 中键入公式“=G5/COUNT(D:D)”，获得第一组平均秩次，将此公式向右拖放至 H9，计算第二组平均秩次。

⑤在单元格区域 H11:I13 中，根据统计量 T、检验水准、n_1、n_2 等，在右侧表格中查找相应临界值，并根据假设检验下结论。其中 H12 单元格的公式是“=IF(OR($I$5<=$I$11,$I$5>=$O$7),"P 值小于等于","P 值大于")”，H13 单元格的公式是“=IF(OR($I$5<=$H$11,$I$5>=$I$11),"拒绝 H0","不拒绝 H0")”，用于判断是否 P 值小于等于 0.05，是否拒绝 H_0。

⑥在单元格 K5 中键入公式“=$G$7*$H$7+$H$7*($H$7+1)/2-$H$5”，在单元格 L5 中键入公式“=$G$7*$H$7+$G$7*($G$7+1)/2-$G$5”，分别计算 U_1、U_2。对于例 9-11，第一组的 10 与第二组所有值比较，等于第二组的 10，大于第二组的 2～9 八个数，所以比较结果是 8.5；第一组的 12 与第二组所有值比较，等于第二组的 12，大于第二组的 2～11 十个数，所以比较结果是 10.5；…。第一组的每一个值轮回与第二组所有值比较得到 8.5+10.5+12×8=115。同理，用第二组的每一个值与第一组的所有值进行比较，将得到 5。

在单元格区域 L7:M9 中，根据 U_1、U_2、n_1、n_2 等，在右侧表格中查找相应临界值，并据此下 Mann-Whitney U 检验结论。

⑦在单元格 L13 中键入公式“=$G$7*($G$7+$H$7+1)/2”，在单元格 L14 中键入公式“=SQRT($G$7*$H$7*($G$7+$H$7+1)/12)”，分别计算近似正态分布均数、标准差。在单元格 L15 中键入公式“=(ABS(I5-L13))/L14”，检验统计量 z 值；在单元格 L16 中键入公式“=(1-NORMSDIST($L$15))*2”，计算 z 值对应的双侧 P 值；在单元格 L17 中键入公式“=(1-NORMSDIST($L$15))”，计算 z 值对应的单侧 P 值。

结果分析：经 Wilcoxon 秩和检验，本例实验组秩和 T_1 为 170，例数 n_1=10，对照组秩和 T_2 为 83，n_2=12，取 n_1 较小的 T_1 为统计量 T，按 n_1=10，n_2-n_1=2 查表，得双侧 0.05 的界值为 84～146，T 落在此范围外，故 P<0.05，拒绝 H_0，可以认为两组小鼠生存日数的差别有统计学意义。Mann-Whitney U 检验和正态近似检验获得了相似的结果。

温馨提示：修改要点如下：

①更换第 A、B 列数据；

②如果必要，可对 H2 的检验水准进行修订。

请注意：本程序的 z 值计算公式没有考虑相同秩次问题，因此，如果相同秩次较多时，获得的结果会与常用统计软件有所不同，建议使用常用统计软件输出的结果。

例 9-12：某研究者选取 32 只同批次的小白鼠进行人参的镇静效果观察。将其中 20 只随机分配到人参组，以 5%人参浸液对其做腹腔注射；12 只分配到对照组，以等量蒸馏水对其做同样注射。实验结果见表 9-8。能否说明人参有镇静作用？（见光盘文件“09-09 两

独立样本非参数检验.xlsx”）

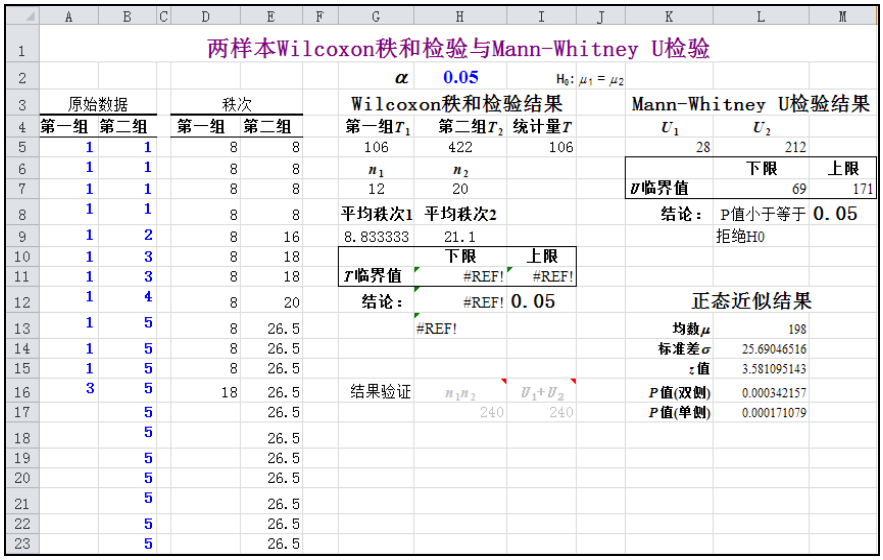
表 9-8　人参的镇静效果观察

镇静等级	人 参 组	对 照 组
−	4	11
±	1	0
+	2	1
++	1	0
+++	12	0

解：很多人看到这种行列表资料就进行卡方检验，实际上这样做是错误的。因为这里的两组观察结果变量是“镇静等级”，是有序的。卡方检验不能利用“有序”的信息。该例正好适合上面介绍的两独立样本的非参数检验。Excel 分析步骤如下：

①整理数据成可用形式。将“镇静等级”“−、±、+、++、+++”数字化为“1、2、3、4、5”。人参组数据实际为 4 个 1、1 个 2、2 个 3、1 个 4 和 12 个 5，而对照组为 11 个 1、1 个 3。

②经 Excel 整理成两列的数据，复制、粘贴（或直接键入）到光盘文件“09-09 两独立样本非参数检验.xlsx”第 A、B 两列的相应位置，便可立即得到如图 9-17 所示的结果。



两样本Wilcoxon秩和检验与Mann-Whitney U检验

						α	0.05	$H_0: \mu_1 = \mu_2$				
	原始数据		秩次			Wilcoxon秩和检验结果			Mann-Whitney U检验结果			
	第一组	第二组	第一组	第二组		第一组T_1	第二组T_2	统计量T	U_1	U_2		
	1	1	8	8		106	422	106	28	212		
	1	1	8	8		n_1	n_2			下限	上限	
	1	1	8	8		12	20		U临界值	69	171	
	1	1	8	8		平均秩次1	平均秩次2		结论：	P值小于等于	0.05	
	1	2	8	16		8.833333	21.1			拒绝H0		
	1	3	8	18			下限	上限				
	1	3	8	18		T临界值	#REF!	#REF!				
	1	4	8	20		结论：	#REF!	0.05	正态近似结果			
	1	5	8	26.5			#REF!		均数μ	198		
	1	5	8	26.5					标准差σ	25.69046516		
	1	5	8	26.5					z值	3.581095143		
	3	5	18	26.5		结果验证	n_1n_2	U_1+U_2	P值(双侧)	0.000342157		
		5		26.5			240	240	P值(单侧)	0.000171079		
		5		26.5								
		5		26.5								
		5		26.5								
		5		26.5								
		5		26.5								
		5		26.5								

图 9-17　例 9-12 分析结果

结果分析：经 Wilcoxon 秩和检验，本例对照组平均秩次为 8.83，而人参组平均秩次为 21.1，相差较大。但由于右侧的 T 界值表 n_1 最多只有 10，所以程序查找不到 T 界值，自

然也就不能下假设检验的结论。Mann-Whitney U 检验允许 n_1、n_2 最多分别为 20。获得 U_1=28<69、U_2=212>171，所以得到 P<0.05，有理由拒绝 H_0，可以认为人参的镇静作用有效。

9.3.3　多样本 Kruskal-Wallis 秩和检验

方差分析需满足独立、正态、方差齐同等基本条件。然而，实际研究中的数据常常不能满足这些条件，如许多数据为偏态（非正态）分布、多个组中有一个组的方差太大或太小、观察指标是有序分类变量等，此时不能使用方差分析，但可采用 Kruskal-Wallis 检验。Kruskal-Wallis 检验是一种类似于 Wilcoxon 秩和检验的方法，可看作 Wilcoxon 秩和检验的拓广，用来检验多个样本所来自的总体分布是否相同。

Kruskal-Wallis 检验的基本思想是：先将各处理组数据混合在一起按从小到大顺序进行编秩，如果有相同数据则取平均秩次，记观测值 Y_{ij} 的秩为 R_{ij}，对每一组观测值的秩求和得到第 i 组秩和 $R_i=\sum_{j=1}^{n_i}R_{ij}$，其中 $i=1,\cdots,k$ 是每组的编号，$j=1,\cdots,n_i$ 是每组内部个体值的编号。由 $\bar{R}_i=R_i/n_i$ 计算每组的平均秩次，如果无效假设（H_0：k 个总体分布位置相同）为真，则秩应该在 k 个组之间均匀分布，每组实际的平均秩 $\bar{R}_i$ 与所有资料的平均秩 $\bar{R}=(N+1)/2$ 的偏差应该很小。若这些 $\bar{R}_i$ 差异很大，则下面公式计算得到的 H 值将较大，就可以拒绝无效假设。K-W 检验的检验统计量为：

$$H=\frac{12}{N(N+1)}\sum_{i=1}^{k}n_i(\bar{R}_i-\bar{R})^2=\frac{12}{N(N+1)}\sum_{i=1}^{k}\frac{R_i^2}{n_i}-3(N+1)$$

其中 $N=\sum_{i=1}^{k}n_i$ 是总例数，k 是待比较的组数，n_i 是第 i 组的样本含量（可不同）。

当存在较多相同秩次时，检验统计量 H 可以修正为：

$$H_c=\frac{H}{C}=\frac{H}{1-\sum(t_j^3-t_j)/(N^3-N)}$$

式中，t_j 为第 j 个相同秩次的个数，C 为校正系数。

当样本含量较大（如每组样本含量大于等于 4）时，统计量 H 近似服从自由度为 k-1 的 χ^2 分布；当样本含量较小时（如每组样本含量小于 4），可查秩和检验 H 界值表（见光盘文件“09-10 Kruskal-Wallis 秩和检验.xlsx”）进行判断。

类似于方差分析，如果拒绝 H_0，还应进行两两差异的多重比较。设 R_A 和 R_B 分别为第 A 组和第 B 组样本的秩和，其平均秩和分别为 $\bar{R}_A$ 和 $\bar{R}_B$。当样本含量较大时，可采用 Wilcoxon 秩和检验的正态近似法计算 z 值，公式为：

$$z=\frac{\bar{R}_A-\bar{R}_B}{\sigma_{\bar{R}_A-\bar{R}_B}}=\frac{\bar{R}_A-\bar{R}_B}{\sqrt{\dfrac{N(N+1)}{12}\left(\dfrac{1}{n_A}+\dfrac{1}{n_B}\right)\left(1-\dfrac{\sum(t_i^3-t_i)}{N^3-N}\right)}}$$

所有符号含义与前面相同。由标准正态分布可得到 z 对应的 P 值。Bonferroni 校正 P 值为 z 检验 P 值的 $k(k-1)/2$ 倍。

例 9-13：对按完全随机设计分组的四组大白鼠，给予不同剂量的某种激素后，测量耻骨间隙宽度的增加量（mm）结果如表 9-9 所示。试分析各组的增加量有无差异？

表 9-9　耻骨间隙宽度的增加量（mm）

1 组	0.15	0.30	0.40	0.50		
2 组	1.20	1.35	1.40	1.50	1.90	2.30
3 组	0.50	1.20	1.40	2.00	2.20	2.20
4 组	1.50	1.50	2.50	2.50		

解：采用后面章节介绍的 QQ 图、PP 图可对数据的正态性进行直观判断，采用第 8 章的 Levene 或 Bartlett 齐性检验可对资料方差进行齐性检验。经检验该数据违背方差分析的基本条件，所以采用 Kruskal-Wallis 检验。由 Excel 实现该资料分析的基本思想如下：

① 如图 9-18 所示（见光盘条件“09-10 Kruskal-Wallis 秩和检验.xlsx”），在第 A、B 列键入原始数据组别（注意必须采用阿拉伯数字）和结果变量数据。在单元格 C5 中键入公式“=IF(AND(A5<>"",B5<>""),COUNTIF($B:$B,"<"&B5)+(1+COUNTIF($B:$B, "="&B5))/2,"")”，对第 B 列数据进行混合编秩，将此公式拖放至第 204 行（即程序可处理的总例数为 200）。

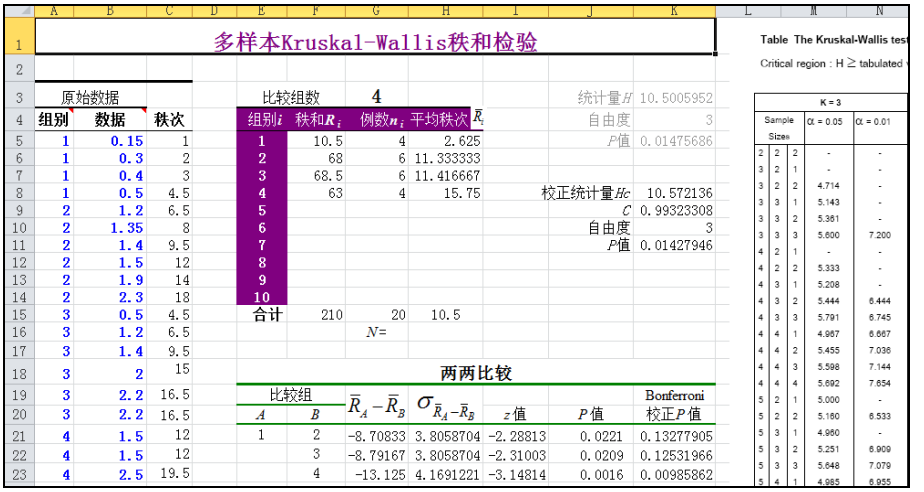


图 9-18　例 9-13 的 Kruskal-Wallis 检验界面

②在单元格 G3 中键入公式“=MAX(A:A)“确定比较组数。在单元格 F5 中键入公式“=IF(E5<=$G$3,SUMIF(A:A,"=1",C:C),"")”计算，将此公式拖放至 F14，并修改其中的“1”为“2～10”，求每组秩和（由此可见，该程序可以设定最多可分析 10 组数据）。同理，可在 G5:G14 中计算每组例数。在单元格 H5 中键入公式“=IF(F5<>"",F5/G5,"")”，将此公式拖放至 H14，计算每组的平均秩次。F15:H15 分别计算了总秩和、总例数和总平均秩次。

③单元格区域 J3:K11 用于计算统计量 H 及相应 P 值。最主要结果校正统计量 H_C 为校正相同秩次后统计量，公式为“=K3/K9”，式中 K3 单元格公式为“=12/G15/(G15+1)*SUM(IF(F5:F14<>"",F5:F14^2/G5:G14,0))-3*(G15+1)”，K9 单元格公式为“=1-SUM(N:N)/(G15^3-G15)”。其中第 N 列为相同秩次的有关计算结果。

④E18:K35 进行了非参数两两比较（类似第 8 章方差分析的两两比较）。

⑤第 L 列以右的表格附有较小样本比较的秩和检验 H 界值表，可供查用。

结果分析：由程序计算得到校正统计量 H_C=10.572，对应自由度=3 的 χ^2 分布 P 值=0.014<0.05，按 α＝0.05 检验水准拒绝 H_0，即可认为各组耻骨间隙宽度的增加量差别具有统计学意义。经两两比较可见，只有第 1 组与第 4 组的 Bonferroni 校正 P＝0.0099 < 0.05，差异有统计学意义，其他组之间的差异没有统计学意义。

温馨提示：更换第 A、B 列数据，便可获得所有结果。注意：在第 A 列中组别输入只允许阿拉伯数字，且第 B 列的第一个数据应是最小值，如果第 B 列第 1 个数字不是最小值，可将最小值对应行数据移至数据顶部。

例 9-14：四种疾病患者痰液内嗜酸性白细胞的检查结果见表 9-10，问四种疾病患者痰液内的嗜酸性白细胞有无差别？

表 9-10　四种疾病患者痰液内的嗜酸性白细胞比较

白细胞数量	四种疾病			
	支气管扩张	肺 水 肿	肺　　癌	病毒性呼吸道感染
–	0	3	5	3
+	2	5	7	5
++	9	5	3	3
+++	6	2	2	0

解：很多人看到这种行列表资料就进行卡方检验，实际上这样做是错误的。因为这里的四组观察结果变量“白细胞数量”是有序的，卡方检验不能利用“有序”的信息。该例正好适合上面介绍的 Kruskal-Wallis 检验。Excel 分析步骤如下：

①整理数据成可用形式。将“白细胞数量”“–、+、++、+++”数字化为“1、2、3、4”。

四种疾病：支气管扩张、肺水肿 、肺癌、病毒性呼吸道感染也分别数字化为“1、2、3、4”。

②清除光盘文件“09-10 Kruskal-Wallis 秩和检验.xlsx”的第 A、B 列的数据，然后将经过 Excel 整理成两列的数据（第 1 列为四种疾病的编号，第 2 列为白细胞等级数据）复制、粘贴（或直接键入）到该文件第 A、B 列相应位置。注意：因为最顶上的数据必须为最小值，该例如果不调整将得不出正确结果。使数字最小值“1 放在顶行数据后，得到如图 9-19 所示的结果（也见光盘文件“09-10 Kruskal-Wallis 秩和检验.xlsx”）。

结果分析：由程序计算得到校正统计量 H_C=15.583，对应自由度=3 的χ^2 分布 P 值=0.0014<0.05，按α＝0.05 检验水准拒绝 H_0，即可认为四种疾病患者痰液内的嗜酸性白细胞差别具有统计学意义。经两两比较可见，支气管扩张组与肺癌、病毒性呼吸道感染的嗜酸性白细胞差别具有统计学意义，其他两两之间差异无统计学意义。

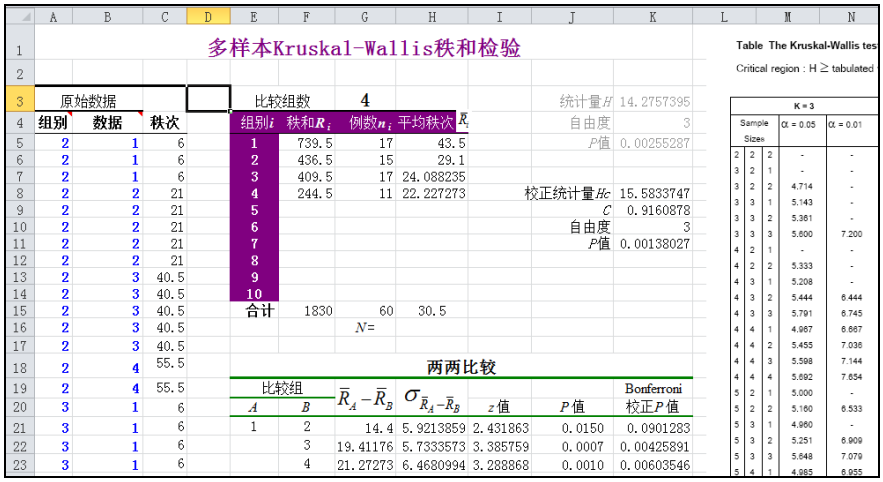


图 9-19　例 9-14 的 Kruskal-Wallis 检验界面

9.3.4　多个相关样本的 Friedman 检验

用于随机区组设计或其他的多个相关样本比较，目的是比较多个总体中位数是否相同。

H_0：多个总体的中位数相同

H_1：多个总体的中位数不同或不全相同

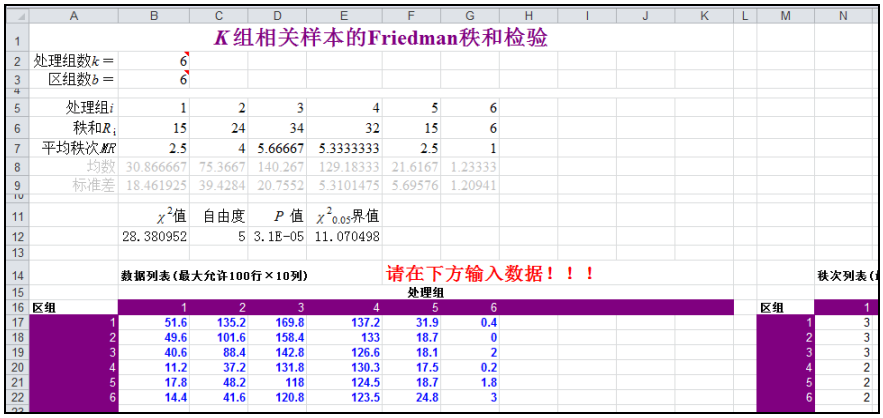
进行 Friedman 检验时，首先在各区组内将观察值按其数值由小到大排列秩次，若有相同秩次则取平均秩次。然后再按处理组求秩和 R_i，通过检验统计量：

$$\chi^2=\frac{12}{bk(k+1)}\sum_{i=1}^{k}R_i^2-3b(k+1)$$

作出统计学推断。其中 b 为区组数，k 为处理组数，R_i 为各处理组的秩和，$i=1,2,\cdots,k$。

该统计量近似服从自由度为（k-1）的 χ^2 分布。

例 9-15：6 条狗服用阿司匹林后不同时间（小时）点的血中药物浓度（μg/ml）如图 9-20 所示，问服药后各时间点血中药物浓度有无差别？



K组相关样本的Friedman秩和检验

处理组数k =	6						
区组数b =	6						
处理组i	1	2	3	4	5	6	
秩和R_i	15	24	34	32	15	6	
平均秩次MR	2.5	4	5.66667	5.3333333	2.5	1	
均数	30.866667	75.3667	140.267	129.18333	21.6167	1.23333	
标准差	18.461925	39.4284	20.7552	5.3101475	5.69576	1.20941	

χ^2值	自由度	P 值	$\chi^2_{0.05}$界值
28.380952	5	3.1E-05	11.070498

数据列表(最大允许100行×10列)　　请在下方输入数据！！！　　秩次列表(

	处理组							
区组	1	2	3	4	5	6	区组	1
1	51.6	135.2	169.8	137.2	31.9	0.4	1	3
2	49.6	101.6	158.4	133	18.7	0	2	3
3	40.6	88.4	142.8	126.6	18.1	2	3	3
4	11.2	37.2	131.8	130.3	17.5	0.2	4	2
5	17.8	48.2	118	124.5	18.7	1.8	5	2
6	14.4	41.6	120.8	123.5	24.8	3	6	2

图 9-20　多个相关样本的 Friedman 检验

解：图 9-20 来自光盘文件“09-11 Friedman 检验.xlsx”。为了使程序能够处理 100 个以内的区组，10 个以内的处理组，且为了能方便看到结果，所以将要输入数据的区域放在了结果区域的下方。Friedman 秩和检验可按下列步骤进行。

① 按图 9-20 将原始数据输入到工作表的单元格区域 B17:G22 中的相应位置。

② 在单元格 B2 中输入公式“=COUNT(B17:K17)”，在 B2 中输入公式“=COUNT(B17:B116)”，分别确定数据的处理组数和区组数。

在单元格 B16 中输入公式“=IF(COLUMN(F14)-5<=行列格式的原数据与秩次!$B$2,COLUMN(F14)-5,"")”，向右拖放至 K16 自动产生处理组序号；在单元格 A17 中输入公式“=IF(ROW(A16)-15<=行列格式的原数据与秩次!$B$3,ROW(A16)-15,"")”，向下拖放至 A116 自动产生区组序号。

③ 在数据区的右侧是对应的按区组排秩次的秩次区域。在单元格 N17 中输入公式“=IF($M17<>"",IF(N$16<>"",COUNTIF($B17:$G17,"<"&B17)+(1+COUNTIF($B17:$G17,"="&B17))/2,""),"")”，然后向右下拖放至 W116，获得每个数据对应的各区组内秩次，并且相同秩次取平均秩次。

④ 在单元格区域 B5:K9，各行分别显示了处理组编号 i、秩和 R_i、平均秩次 MR、算术均数、标准差。均数、标准差仅供参考，所以显示为淡灰色。

⑤ 在单元格 B12 中键入公式“=12/(B3*B2*(B2+1))*SUMSQ(B6:K6)-3*B3*(B2+1)”，获得检验统计量χ^2值；在单元格 C12 中键入公式“=B2-1”，计算χ^2值对应的自由度；在单

元格 D12 中键入公式“=CHIDIST(B12,C12)”，计算χ^2值对应的概率 P 值；在单元格 E12 中键入公式“=CHIINV(0.05,C12)”，计算相应自由度与检验水准α=0.05 的χ^2分布界值。

结果分析：由图 9-20 所示的结果可见，χ^2=28.381，处理组数 k=6，自由度为 5，其对应的 P 值为 3×10^{-5}<0.05，拒绝 H_0，即可以认为服用阿司匹林后不同时间（小时）点血中药物浓度的差别具有统计学意义。

温馨提示：更换单元格区域 B17:G22 中的相应位置的数据，便可获得所有结果。注意：每一行为一个区组，每一列为一个处理组。输入数据后区组、处理组序号会自动显示。

此外，光盘文件“09-11 Friedman 秩和检验 2.xlsx”可以将数据按三列输入，第一列为区组、第二列为处理组、第三列为结果变量数据，如图 9-21 所示，结果显示在数据的右侧。数据的行列表形式及其相应秩次被存放在“行列格式的原数据与秩次”工作表中。

	A	B	C	D	E	F	G	H	I	J	K
1	K组相关样本的Friedman秩和检验										
2											
3			结果变量		处理组数k =	6					
4	区组	处理组	数据		区组数b =	6					
5	1	1	51.6								
6	2	1	49.6		处理组i	1	2	3	4	5	6
7	3	1	40.6		秩和R_i	15	24	34	32	15	6
8	4	1	11.2		平均秩次$\bar{R}_i$	2.5	4	5.66667	5.333333333	2.5	1
9	5	1	17.8		均数	30.86667	75.366667	140.267	129.1833333	21.6167	1.23333
10	6	1	14.4		标准差	18.46192	39.428399	20.7552	5.310147518	5.69576	1.20941
11	1	2	135.2								
12	2	2	101.6								
13	3	2	88.4			χ^2值	自由度	P 值	$\chi^2_{0.05}$界值		
14	4	2	37.2			28.38095	5	3.1E-05	11.07049769		

图 9-21　Friedman 秩和检验的第二个程序界面

9.4　Ridit 分析

“Rid”是英文“relative to an identified distribution”的缩写，“it”是 unit 的字尾，因此，Ridit 分析也称为参照单位分析，可用于两样本或多样本有序分类（等级）资料（如临床疗效为治愈、显效、好转、无效和恶化；检验结果为–、±、+、++等）的比较。

1. 两组比较的 Ridit 分析

此法是先选定一个标准组（常用例数较多的组或取两组的合并），将其分布视作总体分布，计算出其 R 值，再按标准组各等级 R 值计算出对比组的平均 Ridit 值（$\bar{R}_i$），然后进行 z 检验。公式为：

$$\overline{R}=\frac{\sum fR}{n},\quad Z=\frac{\overline{R}_1-\overline{R}_2}{\sqrt{\frac{1}{12}\left(\frac{n_1+n_2}{n_1n_2}\right)}}$$

$\bar{R}_1$、$\bar{R}_2$的 95%置信区间分别为：$\bar{R}_1 \pm 1.96\frac{1}{\sqrt{12n_1}}$、$\bar{R}_2 \pm 1.96\frac{1}{\sqrt{12n_2}}$。

式中，f为对比组各等级频数，n 为对比组总例数，n_1、n_2为对比组每组例数。标准组的平均值$\bar{R}$等于 0.5，而对比组平均 Ridit 值$\bar{R}_1$、$\bar{R}_2$的取值在 0～1 之间。如果对比组的 95%置信区间不包括 0.5，则对比组与标准组差别有统计学意义。

例 9-16：某医院用 A、B 两种配方治疗慢性阑尾炎，结果见表 9-11。试分析两种配方治疗慢性阑尾炎疗效有无差别？(见光盘文件“09-12 2 组 Ridit 分析.xlsx”)

表 9-11　两种配方治疗慢性阑尾炎疗效

组　别	无　效	好　转	显　效	治　愈	合　计
A 配方	8	23	28	33	92
B 配方	12	26	16	24	78

解：两组比较的 Ridit 分析步骤如下：

①第 A、B、C 列为数据区，A 列为等级，本程序设置最多 10 个等级；B、C 列分别为两组各等级对应的频数。

②在单元格 G3、G4 中分别键入公式“=SUM(B:B)”、“=SUM(C:C)”，获得 A、B 组的频数之和。

③单元格区域 F7:F16 中是标准组的各等级频数f，本例标准组频数取两组的合并值，如单元格 F7 中的公式为“=IF(B4<>"",B4+C4,"")”，计算标准组的等级 1。在单元格 F17 中键入公式“=SUM(F7:F16)”，计算标准组频数合计。

④在单元格 G7 中键入公式“=IF(F7<>"",F7/2,"")”，计算等级 1 对应的$f/2$，将此公式向下拖放填充至单元格 G16，求其他等级的$f/2$。

⑤计算 f 的累计和并移下一行。让单元格 H7 空缺，在单元格 H8 中输入公式“=IF(F8<>"",H7+F7,"")”，将此公式向下拖放填充至单元格 H16。

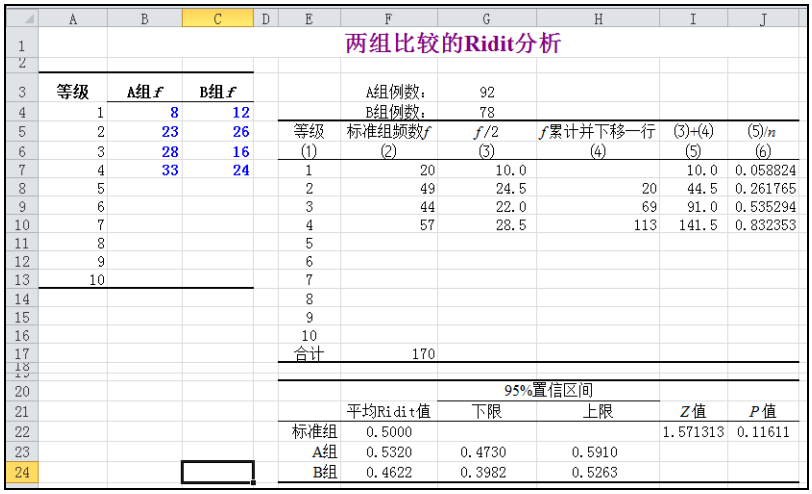
⑥在单元格 I7 中键入公式“=IF(F7<>"",G7+H7,"")”，计算本行前两项之和，将此公式向下拖放填充至单元格 I16。

⑦计算各等级的 Ridit 值。在单元格 J7 中键入公式“=IF(F7<>"",I7/$F$17,"")”，计算等级 1 的 Ridit 值，将此公式向下拖放填充至单元格 J16。

⑧在单元格 F22、F23、F24 中分别键入公式“=SUMPRODUCT(J7:J16,F7:F16)/F17”、“=SUMPRODUCT(B4:B13,J7:J16)/G3”、“=SUMPRODUCT(C4:C13,J7:J16)/G4”，计算标准组、A 组、B 组的平均 Ridit 值。

⑨在单元格 G23、H23 中分别键入公式“=$F$23-1.96/SQRT(12*$G$3)”、“=$F$23+1.96/SQRT(12*$G$3)”，计算 A 组平均 Ridit 值对应的 95%置信区间上、下限值。在单元格 G24、

H24 中分别键入公式“=$F$24-1.96/SQRT(12*$G$4)”、“=$F$24+1.96/SQRT(12*$G$4)”，计算 B 组平均 Ridit 值对应的 95%置信区间上、下限值。



两组比较的Ridit分析

等级	A组f	B组f
1	8	12
2	23	26
3	28	16
4	33	24
5		
6		
7		
8		
9		
10		

A组例数：92
B组例数：78

等级	标准组频数f	f/2	f累计并下移一行	(3)+(4)	(5)/n
(1)	(2)	(3)	(4)	(5)	(6)
1	20	10.0		10.0	0.058824
2	49	24.5	20	44.5	0.261765
3	44	22.0	69	91.0	0.535294
4	57	28.5	113	141.5	0.832353
5					
6					
7					
8					
9					
10					
合计	170				

		95%置信区间			
	平均Ridit值	下限	上限	Z值	P值
标准组	0.5000			1.571313	0.11611
A组	0.5320	0.4730	0.5910		
B组	0.4622	0.3982	0.5263		

图 9-22　两组比较的 Ridit 分析

⑩在单元格 I22 中键入公式“=(F23-F24)/SQRT((G3+G4)/(12*G3*G4))”，计算两组比较的 z 值。在单元格 J22 中键入公式“=(1-NORMSDIST(ABS(I22)))*2”，计算 z 值对应的标准正态分布概率，即 P 值（双侧）。

结果分析：由图 9-22 可见，本例 A 组的平均 Ridit 值及 95%置信区间为 0.5320（0.4730，0.5910），B 组为 0.4622（0.3982，0.5263），均包括 0.5，故差别无统计学意义。由统计量 z=1.5713，对应 P 值为 0.1161>0.05，不拒绝 H_0，其结论与置信区间相同。

温馨提示：更换光盘文件“09-12 2 组 Ridit 分析.xlsx”单元格区域 B4:C13 中的相应位置的数据，便可获得所有结果。注意：A 列为等级序号，B 列为第 1 个比较组频数，C 列为第 2 个比较组频数。在置信区间信息后的 Ridit 图形可供结果判断参考。

2. 多组比较的 Ridit 分析

多组比较的 Ridit 分析中标准组和对比组的平均 Ridit 值计算方法同上，不同的是假设检验用 χ^2 检验。公式为：

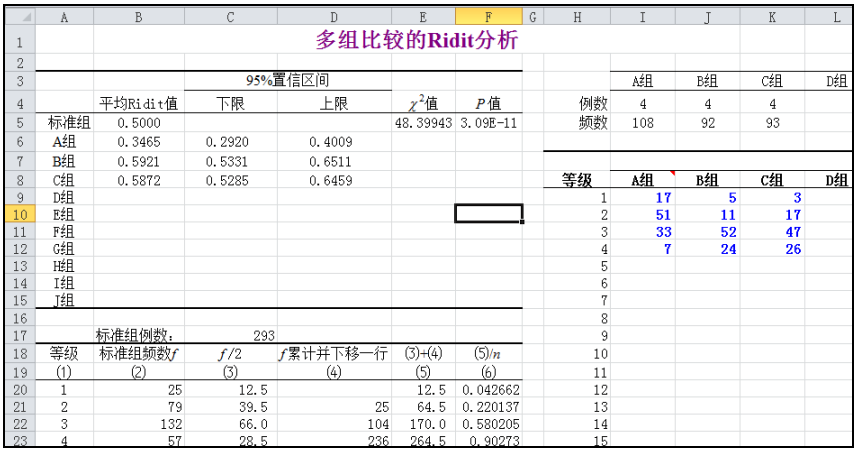
$$\chi^2 = 12\sum n_i(\overline{R}_i - 0.5)^2$$

此时，χ^2 统计量近似服从自由度=（对比组数–1）的 χ^2 分布。

例 9-17：用 A、B、C 三种药物治疗单纯性慢性气管炎，其疗效如表 9-12 所示。试分析三种药物疗效有无差别？（见光盘文件“09-13 k 组 Ridit 分析.xlsx”）

表 9-12　三种配方治疗单纯性慢性气管炎疗效

组　别	治　愈	显　效	好　转	无　效	合　计
A 药	17	51	33	7	108
B 药	5	11	52	24	92
C 药	3	17	47	26	93
合计	25	79	132	57	293



多组比较的Ridit分析

		95%置信区间			
	平均Ridit值	下限	上限	χ^2值	P值
标准组	0.5000			48.39943	3.09E-11
A组	0.3465	0.2920	0.4009		
B组	0.5921	0.5331	0.6511		
C组	0.5872	0.5285	0.6459		
D组					
E组					
F组					
G组					
H组					
I组					
J组					

	标准组例数：	293			
等级	标准组频数f	$f/2$	f累计并下移一行	(3)+(4)	(5)/n
(1)	(2)	(3)	(4)	(5)	(6)
1	25	12.5		12.5	0.042662
2	79	39.5	25	64.5	0.220137
3	132	66.0	104	170.0	0.580205
4	57	28.5	236	264.5	0.90273

	A组	B组	C组	D组
例数	4	4	4	
频数	108	92	93	

等级	A组	B组	C组	D组
1	17	5	3	
2	51	11	17	
3	33	52	47	
4	7	24	26	

图 9-23　多组比较的 Ridit 分析

解：多组比较的 Ridit 分析可按下列步骤进行。

①该程序可分析 50 个等级，比较 2～10 个组间是否有差异。当只有两个比较组时，获得的检验 P 值相同，$\sqrt{\chi^2}=z$，如用例 9-16 数据更新该程序数据，获得的 $\sqrt{\chi^2}=\sqrt{2.469}=1.571=z$。

②右下方是数据输入区域。本例数据存放在 I9:K12。

③单元格区域 B18:B69、C18:C69、D18:D69、E18:E69、F18:F69 分别为标准组频数、$f/2$、f 累计下移一行、(3)+(4)、(5)/n；B3:D15 为各组平均 Ridit 值和 95%置信区间，上述计算与两组比较的 Ridit 分析相同。

④在单元格 E5 中键入公式“=12*SUM(G6:G15)”，计算多组比较的χ^2值。其中 G6 的计算公式为“=IF(B6<>"",$I$59*(B6-0.5)^2,"")”，用此公式计算 A 组的 $n_A(\bar{R}_A-0.5)^2$。G7:G15 完成相同的计算。

⑤在单元格 F5 中键入公式“=CHIDIST(E5,COUNT(B6:B15)-1)”，计算χ^2值在自由度为（组数–1）时对应的χ^2分布概率，即 P 值。

⑥单元格区域 I4:R4、I5:R5 分别存放比较组的例数与频数。

结果分析：由图 9-23 可见，三个用药组的 Ridit 均值及 95%置信区间分别为 0.3465（0.2920，0.4009）、0.5921（0.5331，0.6511）、0.5872（0.5285，0.6459），95%置信区间均不包括 0.5，故差别有统计学意义。统计量χ^2=48.3994，自由度=3，对应P值为3.09×10^{-11}<0.05，拒绝H_0，其结论与置信区间相同。

另外，从置信区间来看，A 组的 95%置信区间不与 B、C 组的 95%置信区间相交，而 B、C 组间互有交叉，说明 A 组与 B、C 组的差别有统计学意义，A 组疗效低于 B、C 组；B、C 组间的差别无统计学意义。

在该文件右侧有一个 Ridit 分析图（如图 9-24），可直观地看出比较组之间的差异。

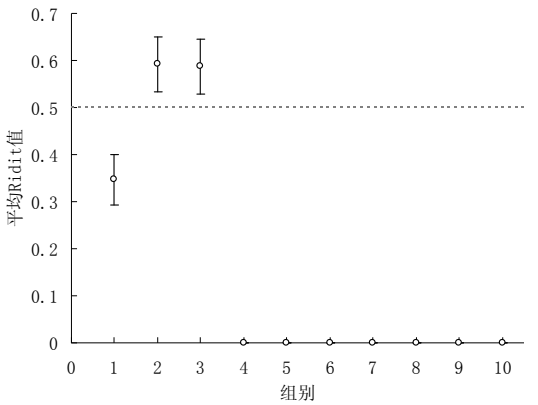


图 9-24　Ridit 分析图

温馨提示：更换光盘文件“09-13 k 组 Ridit 分析.xlsx”单元格区域 I9:R58 中的相应位置的数据，便可获得所有结果。

无论是两组比较，还是多组比较，本程序均采用了各组频数之和作为标准组。

关于对比组平均 Ridit 值>0.5 或<0.5 的具体含义，应以标准组频数的累计方向而定。以上两个例子的累计方向均为“无效→治愈”，则疗效最好的“治愈”等级 Ridit 值必然最大；此时，对比组平均 Ridit 值>0.5 就表示疗效优于标准组，<0.5 表示疗效不如标准组。相反，标准组若按“治愈→无效”的方向累计，则疗效最差的“无效”等级 Ridit 值将会最大；此时，对比组平均 Ridit 值>0.5 反倒表示疗效不如标准组，而<0.5 却表示优于标准组。

第 10 章

相关与回归

为了反映两个或多个变量之间的相互关系，描述相关关系的方向与密切程度，需要采用相关分析；为了反映两个或多个变量之间的依存关系，建立回归方程，需要采用回归分析。可用于相关与回归的 Excel 数据分析工具有：相关系数、协方差、回归等；此外，在图表中添加趋势线，利用统计函数编制 Excel“程序”，也可用于相关与回归分析。本章将按简单相关与回归、非线性回归、多重相关与回归的先后顺序进行介绍。

10.1 简单相关与回归

简单相关与回归（simple linear correlation and regression）用于进行两个变量之间的相关与回归分析，如分析年龄与身高、肺活量与体重、药物剂量与动物死亡率等之间的关系。例 10-1 的目的是分析“汽车流量”变量是否与“大气二氧化氮浓度”有关系。

例 10-1：某市环境监测站测定的某交通点 10 个时间（$n=10$）的汽车流量与大气二氧化氮（NO_2）浓度数据如下，试进行相关与回归分析。

汽车流量（辆/小时）X	1300	1444	1652	1756	1200	1500	1476	1820	1060
NO_2（mg/L）Y	0.066	0.076	0.170	0.156	0.040	0.120	0.129	0.135	0.029

进行两个变量的相关与回归分析前，应绘制散点图（scatter plot），从图中考察两个变量之间是否呈直线关系，如果散点呈直线趋势，如图 10-1（左）所示，则按本节进行分析；如果散点呈非线性（非直线）趋势，如图 10-1（右）所示，则按下一节非线性回归进行分析。

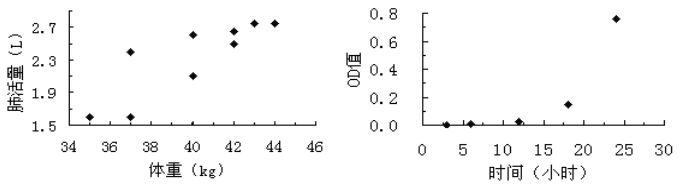


图 10-1　线性趋势（左）　指数曲线趋势（右）散点图

简单线性相关与回归只考虑两个变量之间的直线关系。简单线性相关（simple linear correlation）用于双变量（X 与 Y）正态分布（bivariate normal distribution）资料，通常用 Pearson 相关系数（r, correlation coefficient）来说明两个变量间线性相关关系的密切程度（r 的绝对值大小）和方向（r 的正与负），Pearson 相关系数 r 的数学计算公式为：

$$r=\frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum(X-\bar{X})^2}\sqrt{\sum(Y-\bar{Y})^2}}=\frac{l_{XY}}{\sqrt{l_{XX}}\sqrt{l_{YY}}}$$
$$=\frac{l_{XY}/(n-1)}{\sqrt{l_{XX}/(n-1)}\sqrt{l_{YY}/(n-1)}}=\frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{cov}(X,X)}\sqrt{\mathrm{cov}(Y,Y)}} \quad (10\text{-}1)$$

其中，l_{XY} 表示 X 与 Y 的离均差积和（交叉乘积和），l_{XX}、l_{YY} 分别表示 X、Y 的离均差平方和；$\mathrm{cov}(X,Y)$ 为 X 与 Y 的协方差，$\mathrm{cov}(X,X)$、$\mathrm{cov}(Y,Y)$ 分别表示 X、Y 的方差，可采用协方差工具计算（见 10.1.2 节）。n 为观察个体数。相关系数 r 没有单位，其绝对值$|r|\leqslant 1$。$r>0$，说明变量 X 和 Y 之间为正相关关系，即 X 和 Y 的变化趋势方向相同；$r<0$，则说明变

量 X 和 Y 为负相关关系，表明 X 和 Y 之间反方向变化，如 X 增大 Y 减小。$|r|$越接近于 1，线性关系越密切；$|r|=1$ 为完全相关；$|r|=0$ 则 X 和 Y 之间无线性关系（但有可能有曲线关系）。

检验假设 H_0：总体相关系数$\rho=0$

H_1：总体相关系数$\rho\neq0$

检验统计量：$t_r=\dfrac{r-0}{s_r}$，r 的标准误 $S_r=\sqrt{\dfrac{1-r^2}{n-2}}$，自由度 $\nu=n-2$ （10-2）

简单相关描述了两个随机变量之间线性关联的程度，变量间无主次之分。为了进一步研究其中一个变量如何定量地影响另一个变量，需要采用简单线性回归或非线性回归。

采用直线回归方程反映应变量（dependent variable）Y 依赖于自变量（independent variable）X 的数量关系。表达式为：

$$\hat{Y}=a+bX \tag{10-3}$$

式中，a 为回归直线在 Y 轴上的截距（intercept），$a=0$、$a>0$ 和 $a<0$ 分别表示回归直线与纵轴的交点在原点、原点之上和原点之下。b 为回归系数（regression coefficient），即直线的斜率（slope），表示变量 X 每增加（或减少）一个单位，Y 平均改变 b 个单位。

根据最小二乘法（least square method）原理获得 b，有

$$b=\frac{\sum\left(X-\bar{X}\right)\left(Y-\bar{Y}\right)}{\sum\left(X-\bar{X}\right)^2}=\frac{l_{XY}}{l_{XX}},\quad a=\bar{Y}-b\bar{X} \tag{10-4}$$

其中，$\bar{X}$、$\bar{Y}$ 分别为自变量 X 与应变量 Y 的均数，回归直线一定通过（$\bar{X}$，$\bar{Y}$）这一坐标点。

剩余标准差（residual standard deviation）和决定系数（coefficient of determination）可反映模型的拟合情况。剩余标准差越小，表示回归模型估计精度越高，其计算公式为：

$$S_{Y.X}=\sqrt{\sum(Y-\hat{Y})^2/(n-2)} \tag{10-5}$$

决定系数记为 R^2，即 $R^2=\dfrac{SS_{回}}{SS_{总}}=\dfrac{bl_{XY}}{l_{YY}}=\dfrac{l_{XY}^2}{l_{XX}l_{YY}}$，$R^2$ 无单位，取值在 0 到 1 之间，反映了在 Y 的总变异中回归所能解释的比例，也就是回归贡献的相对大小。R^2 越接近 1（或 100%），回归的效果越好。

对样本回归系数 b 是否来自总体回归系数$\beta=0$ 的总体，可采用

$$t_b=\frac{b}{S_b}=\frac{b}{S_{Y.X}/\sqrt{l_{XX}}} \tag{10-6}$$

检验。对于同一资料，t_b 等于相关分析获得的 t_r。

对样本截距 a 是否来自总体截距$\alpha=0$ 的总体，可采用

$$t_a=\frac{a}{S_a},\quad S_a=\frac{S_{Y.X}}{\sqrt{l_{XX}}}\sqrt{\frac{\sum X^2}{n}}=S_b\sqrt{\frac{\sum X^2}{n}} \tag{10-7}$$

检验。

a、b 总体参数 α 、β 的 95%置信区间的计算公式分别为：

$$a\pm t_{0.05/2,(n-2)}S_a\text{、}\quad b\pm t_{0.05/2,(n-2)}S_b \tag{10-8}$$

其中 $t_{0.05/2,(n-2)}$ 是检验水准为 0.05、自由度为 $n-2$ 的双侧 t 界值。S_a 、S_b 分别为 a、b 的标准误，计算公式如上所示。

因为 r 对应总体不像 a、b 近似服从 t 分布，所以为了得到 r 总体参数 ρ 的 95%置信区间，必须事先对 r 作 Z 变换，即 ρ 的 $100(1-\alpha)$%置信区间为：

$$\tanh\left(\tanh^{-1}r\pm Z_{\alpha/2}\Big/\sqrt{n-3}\right) \tag{10-9}$$

式中，tanh 为双曲（hyperbolic）正切函数（$r=\tanh Z=(\mathrm{e}^{2Z}-1)\big/(\mathrm{e}^{2Z}+1)$）；$\tanh^{-1}$ 为反双曲正切函数（$Z=\tanh^{-1}r=0.5\ln[(1+r)/(1-r)]$）。$r$ 的取值范围为-1<r<1，$Z_{\alpha/2}$ 为检验水准=α 的标准正态分布双侧界值，α =0.05 时，$Z_{0.05/2}=1.96$。

直线总体均数 $\mu_{Y|X}$ 的 95%置信区间为

$$\hat{Y}\pm t_{\alpha/2,(n-2)}S_{\hat{Y}}\text{，其中 }S_{\hat{Y}}=S_{Y.X}\sqrt{\frac{1}{n}+\frac{(x_0-\bar{X})^2}{\sum(X-\bar{X})^2}} \tag{10-10}$$

式中，x_0 为人为给定的值，$S_{Y.X}$ 为剩余标准差，n 为观察个体数，X 、$\bar{X}$ 分别为变量 X 的个体值、均数。

直线个体值 Y 的 95%置信区间为：

$$\hat{Y}\pm t_{\alpha/2,(n-2)}S_Y,\quad S_Y=S_{Y.X}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{X})^2}{\sum(X-\bar{X})^2}} \tag{10-11}$$

回归分析的方差分析表计算公式见表 10-1。

表 10-1　方差分析表计算公式

变异来源	*SS*（sum of squares）	*df*	*MS*（mean square）	*F*
回归（regression）	$\sum(\hat{Y}-\bar{Y})^2=b^2l_{XX}$	1	$SS_{回}/\nu_{回}$	$MS_{回}/MS_{剩}$
剩余（residual）	$\sum(Y-\hat{Y})^2=l_{YY}-b^2l_{XX}$	$n-2$	$SS_{剩}/\nu_{剩}$	
总变异（total）	$\sum(Y-\bar{Y})^2=l_{YY}$	$n-1$		

注：$l_{XX}=\sum\left(X-\bar{X}\right)^2$

10.1.1　Excel 函数实现简单相关与回归

当 Y 变量是随机独立变量、且近似服从正态分布要求时，可采用参数法；不满足时应

考虑采用非参数方法，如采用 Spearman 等级相关分析。

1. 参数法

利用 Excel 函数和绘图工具制作的简单相关与回归分析工作表“程序”，在同一个工作表界面，可实现简单相关分析、简单回归分析，并可显示两变量散点图、残差图、直线均数与个体值的置信区间图。

例 10-2：某地方病研究所调查了 8 名正常儿童的年龄与尿肌酐含量（mmol/24h）结果见表 10-2（也见光盘文件“10-1 简单相关与回归.xlsx”）。请分析尿肌酐含量与年龄之间的相关关系，并建立由年龄（X）推算尿肌酐含量（Y）的回归方程。

解：Excel 函数制作直线相关与回归分析的工作表“程序”分成下面几部分。

①该程序设定最多可分析 2001 个观察个体。第 A 列可根据第 B 列是否为数字自动编号，在 A3 单元格键入 1，在 A4 单元格键入公式“=IF(ISNUMBER(B4),A3+1,"")”，将此公式拖放至 A2003。单元格区域 B3:C2003 为用户键入数据区。在第 B 列必须键入自变量 X，在第 C 列必须键入应变量 Y（见图 10-2a）。

表 10-2 8 名正常儿童的年龄与尿肌酐含量

学生编号	1	2	3	4	5	6	7	8
年龄（岁）X	13	11	9	6	8	10	12	7
尿肌酐含量（mmol/24h）Y	3.54	3.01	3.09	2.48	2.56	3.36	3.18	2.65

注：资料来自孙振球主编，《医学统计学》第二版，184 页。

②图 10-2b 为主要结果区域，计算数学公式已在前面给出。单元格区域 E3:F7 给出了 Pearson 相关系数 r 及其假设检验结果。单元格区域 H3:I12 给出了截距 a 与回归系数 b 的值及其假设检验结果，因为 $t_b = t_r$，所以为节约空间没有在程序中显示 t_b 及其对应 P（与相关分析完全相同）。单元格区域 K3:M9 给出了相关系数 r、回归系数 b、截距 a、预测直线个体值、预测直线均数的置信区间。置信区间的置信度可在 K3 中修改。



	A	B	C
1			
2	编号	X	Y
3	1	13	3.54
4	2	11	3.01
5	3	9	3.09
6	4	6	2.48
7	5	8	2.56
8	6	10	3.36
9	7	12	3.18
10	8	7	2.65

图 10-2a 简单相关与回归分析界面

③预测直线个体值、预测直线均数的置信区间分别受相应标准误 S_Y、$S_{\hat{Y}}$（见单元格 F11、F12）的影响，而这两个标准误直接受给定值 x_0（见单元格 F15）的影响。不同的 x_0 会有不同的预测直线个体值、预测直线均数的置信区间。

④单元格区域 E14:F16 与单元格区域 H14:J16，分别在给定一个变量值的情况下根据回归方程推导出另一个变量的值，从而分别实现预测与控制。给定自变量 $X = x_0$ 推导应变量 Y 的值称为预测；

给定应变量 $Y=y$ 推导自变量 X 的值称为控制。

⑤单元格区域 E14:F16 根据表 10-1“方差分析表计算公式”获得直线的方差分析检验结果，该结果与 t_r、t_b 完全等价，即 $\sqrt{F}=t_r=t_b$，三者 P 值相同（见图 10-2b）。

简单相关与回归

编号	X	Y
1	13	3.54
2	11	3.01
3	9	3.09
4	6	2.48
5	8	2.56
6	10	3.36
7	12	3.18
8	7	2.65

相关分析	
Pearson相关系数	0.8818
自由度	6
t_r值	4.5791
P值（双侧）	0.0038
检验水准	0.05
t临界值	2.4469
S_y	0.2094591
$S_{\hat{y}}$	0.0712744

回归分析	
截距a	1.661666667
回归系数b	0.139166667
剩余标准差$S_{Y.X}$	0.196959528
决定系数R^2	0.777517574
S_b	0.030391515
S_a	0.296998349
t_b值	4.5791
t_a值	5.5949
截距a的P值	0.001387405

95%置信区间(confidence interval,CI)	下限	上限
相关系数r	0.467716673	0.978464582
回归系数b	0.064801309	0.213532025
截距a	0.934937887	2.388395446
个体Y值	2.540805473	3.565861194
总体均数$\mu_{Y\|X}$	2.87893111	3.227735557

预测	
请给出x_0=	10
Y的估计值$\hat{Y}$	3.0533333

控制	
请给出y=	2.5
x的估计值	6.023952096

方差分析表（ANOVA table）

变异来源	SS	df	MS	F	P
回归(regression)	0.813429167	1	0.81343	20.9684	0.0038
剩余(residual)	0.232758333	6	0.03879		
总变异(total)	1.0461875	7			

图 10-2b　简单相关与回归分析界面

⑥图 10-2c 的左侧是由实际数据获得的散点图及由最小二乘法原理获得的直线回归方程，由此图可观察实际数据的线性趋势。右侧是残差图，横坐标与左侧图相同，即变量 X；纵坐标为残差（见图 10-2d 的第 R 列），由此图可观察残差散点的分布情况。残差分布应该是围绕 $y=0$ 的 X 轴呈均匀分布。如果呈现其他分布，则表示不适合采用直线回归拟合。

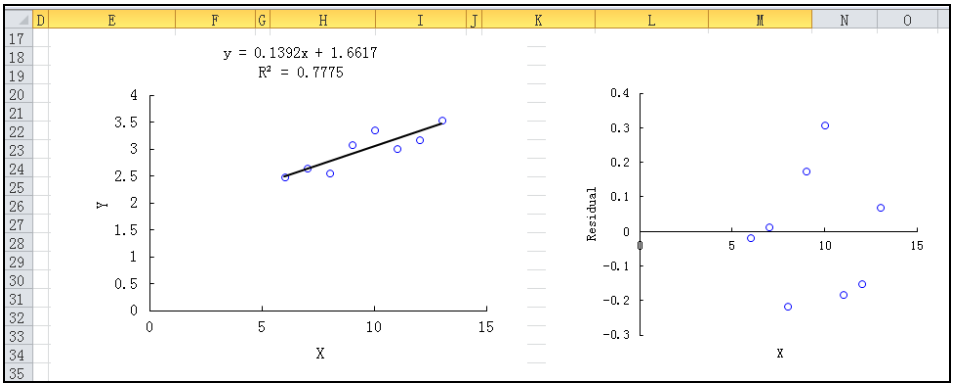


图 10-2c　简单相关与回归分析界面

⑦图 10-2d 主要用于绘制预测直线个体值、预测直线均数的置信区间图，第 U、V、X、Y 列为绘制置信区间图的 Y 轴数据（X 轴数据为用户输入的实际数据 X 值），这些数据受给定值 x_0 大小的影响，不同的 x_0 会有不同的标准误 S_Y、$S_{\hat{y}}$。

结果分析：由图 10-2c 中的散点图可见，散点呈直线趋势。相关分析获得 Pearson 相关系数为 0.8818，其总体相关系数ρ的 95%置信区间为（0.4677，0.9785）。说明年龄与尿肌酐含量间有较大的相关性，经检验 $t_r=4.5791$，相应的 P=0.0038<0.05，在检验水准 α 为 0.05

时，该相关系数具有统计学意义。由回归分析可建立回归方程为：$\hat{Y}=1.6617+0.1392X$，其剩余标准差为 $S_{Y.X}=0.1970$，决定系数 $R^2=0.7775$，表示由直线回归方程可解释 77.75%的 Y 变异信息。回归系数估计值 $b=0.1392$，总体回归系数 β 的 95%置信区间为（0.0648，0.2135），表示在年龄 X（6～13 岁之间）每增加一岁，尿肌酐含量平均增加 0.1392 mmol/24h。

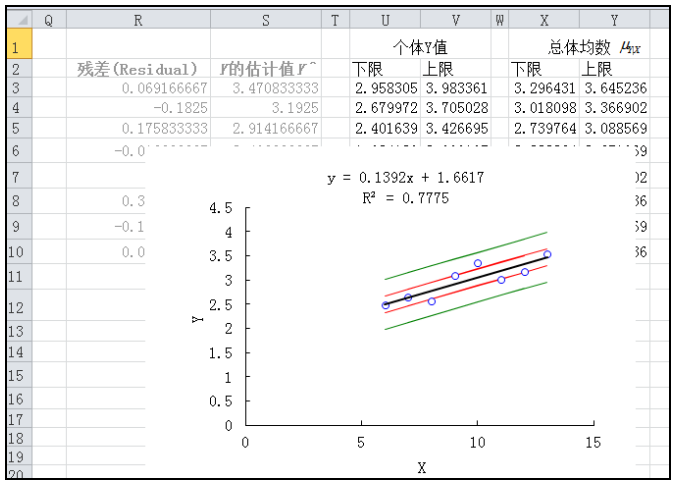


图 10-2d 简单相关与回归分析界面

温馨提示：用户只需以复制、粘贴方式或键盘键入方式，将自己资料的自变量 X 数据放在第 B 列，应变量 Y 数据放在第 C 列，“简单相关与回归”工作表就会自动完成所有计算，包括绘制散点图、残差图、置信区间图。（见光盘文件“10-1 简单相关与回归.xlsx”）

注意：数据应从第 4 行开始输入，程序设计最多可容纳 2001 个 X、Y 对子。

此外用户必须注意修改单元格 F15、I15 的 x_0、y 值。如果需要，单元格 K3 的置信度也可以修改。

2. 秩相关与秩回归

采用 Pearson 简单相关分析，需要两个变量分别是随机、独立（个体间不相关）的正态变量。如果任一个变量不服从正态分布，或者总体分布类型未知以及原始数据用等级表示，则需要采用 Spearman 秩相关系数（Spearman's rank correlation coefficient）来说明两个变量间相关关系的密切程度与方向。

计算原理是：分别将 X 与 Y 从小到大编秩，用所编秩次代替原始数据计算 Pearson 相关系数，并进行相应的假设检验。如果 X 与 Y 对应的秩次为 P 与 Q，则 Spearman 秩相关系数及其检验公式为：

$$r_s=\frac{\sum(P-\bar{P})(Q-\bar{Q})}{\sqrt{\sum(P-\bar{P})^2}\sqrt{\sum(Q-\bar{Q})^2}},\quad t_{r_s}=\frac{r_s}{\sqrt{\left(1-r_s^2\right)/(n-2)}},\quad \nu=n-2 \qquad (10\text{-}12)$$

采用上述简单线性回归的条件是应变量 Y 是随机、独立（个体间不相关）的正态变量，对自变量 X 无特殊要求，如果数据应变量 Y 不满足要求，则可采用秩回归方法。秩回归原理是，先分别对 X 与 Y 从小到大编秩，采用秩次 P 与 Q 建立回归方程，方程公式为：

$$\hat{Q}=a_m+b_mP，\text{其中 } b_m=\frac{\sum(P-\bar{P})(Q-\bar{Q})}{\sum(P-\bar{P})^2},\ a_m=\bar{Q}-b_m\bar{P} \qquad (10\text{-}13)$$

假设检验方法与上述简单线性回归相同。

为了在给定 X 值（X_k）条件下预测相应 Y 的值，需要按以下三步进行。

①用内插法公式将 X_k 值转换为 P_k 值，公式为：

$$P_k=P_i+\frac{(X_k-X_i)}{(X_j-X_i)}(P_j-P_i) \qquad (10\text{-}14)$$

式中，X_k 为给定值（应在样本数据范围内），位于样本中相邻的 X_i 与 X_j 之间（$j>i$），P_i 与 P_j 分别为 X_i 与 X_j 的秩次。

②由公式 $\hat{Q}_k=a_m+b_mP_k$ 计算相应 Y 的秩估计值 $\hat{Q}_k$。

③再次利用内插法求出 $\hat{Q}_k$ 对应的 $\hat{Y}_k$ 值，公式为：

$$\hat{Y}_k=Y_i+\frac{(\hat{Q}_k-Q_i)}{(Q_j-Q_i)}(Y_j-Y_i) \qquad (10\text{-}15)$$

式中，$\hat{Q}_k$ 位于相邻的 Q_i 与 Q_j 之间（$j>i$），Q_i 与 Q_j 分别为 Y_i 与 Y_j 的秩次。

例 10-3：某省调查了 1995 年到 1999 年当地居民 18 类死因的构成，以及每种死因导致的潜在工作损失年数 WYPLL 的构成，结果见图 10-3 第 B、C 列（也见光盘文件“10-2 秩相关与秩回归.xlsx”）。其中 X 为死因构成（%），Y 为 WYPLL 构成（%），试对此作秩相关与秩回归分析（注：资料来自孙振球主编，《医学统计学》第二版）。

解：第 A～C 列与前面叙述相同，即 A 为自动编码。第 B、C 列分别为变量 X、Y 的数据区域，需要用户自我更新。第 E 列为变量 X 的秩次，第 F 列为变量 Y 的秩次。第 G 列为在给定第 E 列 P 值条件下，由求解获得的秩回归方程（$\hat{Q}=0.9020+0.9051P$）估计的 Y 秩次 Q 的估计（预测）值。

主要结果位于单元格区域 J3:N19。单元格区域 J3:K9 为秩相关系数及对此进行假设检验计算结果；单元格区域 M3:N9 用于估计秩回归方程，并对回归系数进行假设检验；单元格区域 J12:K14 用于预测 Y 变量对应的 Q 值（在给定 X 变量对应的秩次 P 条件下）；单元格区域 J16:N19 用于内插法计算，内插法是秩回归预测的中间结果。



秩相关与秩回归

编号	X	Y	X秩次P	Y秩次Q	估计值 $\hat{Q}$
1	0.03	0.05	1	1	1.807018
2	0.14	0.34	2	2	2.712074
3	0.2	0.93	3	6	3.617131
4	0.43	0.69	4	4	4.522188
5	0.44	0.38	5	3	5.427245
6	0.45	0.79	6	5	6.332301
7	0.47	1.19	7	8	7.237358
8	0.65	4.74	8	12	8.142415
9	0.95	2.31	9	9	9.047472
10	0.96	5.95	10	14	9.952528
11	2.44	1.11	11	7	10.85759
12	2.69	3.53	12	11	11.76264
13	3.07	3.48	13	10	12.6677
14	7.78	5.65	14	13	13.57276
15	9.82	33.95	15	18	14.47781
16	18.93	17.16	16	17	15.38287
17	22.59	8.42	17	15	16.28793
18	27.96	9.33	18	16	17.19298

秩相关		秩回归	
Spearman相关系数	0.9051	截距a_m	0.901961
自由度	16	回归系数b_m	0.905057
t_r值	8.5124	剩余标准差$S_{Y.X}$	2.340307
P值（双侧）	0.0000	决定系数R^2	1
检验水准	0.05	S_{bm}	0.106323
t临界值	2.1199	t_{bm}值	8.5124

预测					
请给出秩次P=	7.1667				
Q的估计$\hat{Q}$	7.38823106				
内插法求解					
X_k或Q_k	X_i或Q_i	X_j或Q_j	P_i或Y_i	P_j或Y_j	
0.5	0.47	0.65	7	8	
	P_k或Y_k=	7.166667			

图 10-3 秩相关与秩回归

为了估计当某种死因的构成比为 0.5%时，相应潜在工作损失年数 WYPLL 的构成比是多大？先由 X_k =0.5 在源数据（第 B 列）中找到相邻的 X_i =0.47 与 X_j =0.65，对应的 P_i=7 与 P_j =8，使用内插法求得 P_k =7.1667；然后将 P_k 代入 $\hat{Q}=0.9020+0.9051P$ 中得到 $\hat{Q}_k$ =7.3882；$\hat{Q}_k$ =7.3882 的相邻秩次为 Q_i=7 与 Q_j =8，由此在源数据（第 C 列）中找到相应 Y_i =1.11 与 Y_j =1.19，由内插法得到 $\hat{Y}_k$ =1.14。即当某病的死因构成比为 0.5%时，相应潜在工作损失年数 WYPLL 的构成比是 1.14%。

结果分析：由图 10-3 可见，例 10-3 对应的 Spearman 秩相关系数为 0.9051，经检验概率 P 值=0.0000 远小于 0.05，有统计学意义，即死因构成比对潜在工作损失年数 WYPLL 构成比有影响。秩回归方程为 $\hat{Q}=0.9020+0.9051P$ 。

温馨提示：根据光盘文件“10-2 秩相关与秩回归.xlsx”，用户可将数据输入到 B4:C2003 范围内，便可获得秩相关系数与秩回归方程的有关结果。为了完成预测，还需修改单元格 K13、单元格区域 J18:N18 中的值，所依据的公式是式（10-13）至式（10-15）。

10.1.2 数据分析工具实现简单相关与回归

Excel 数据分析工具（详见 7.1 节）中的“相关系数”可完成 Pearson 相关系数的计算，“回归”可进行简单线性回归分析，“协方差”可完成协方差的计算。这 3 个分析工具对于两个以上变量同样有效，因此，在 10.3 节还要利用这些分析工具。本节仅利用例 10-2 中两个变量数据作相关与回归分析。

1. 相关系数

对于例 10-2 数据，利用 Excel“相关系数”分析工具获得相关系数（correlation coefficient）的步骤如下：

①先将观察数据输入工作表中（见图 10-2a，也见光盘文件“10-3 相关与回归工具.xlsx”）。

②单击的“数据→分析→数据分析→相关系数”，单击“确定”，弹出“相关系数”对话框（见图 10-4），按图 10-4 设置对话框。

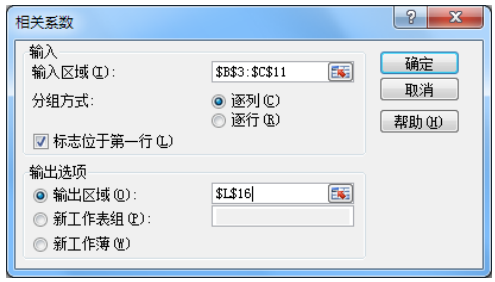


图 10-4 “相关系数”对话框

③单击“确定”，获得变量 X 与 Y 间的相关系数为 0.88177，与上述 Excel 函数实现直线相关与回归分析结果相同。“相关系数”分析工具以“下三角阵”的形式给出相关系数 r（见光盘文件“10-3 相关与回归工具.xlsx”）。

2. 协方差

协方差（covariance）也可以度量两个变量之间的关系，对于 n 对随机样本，X 和 Y 的总体协方差定义为 $\mathrm{cov}(X,Y)=\dfrac{\sum(X-\bar{X})(Y-\bar{Y})}{n}$。一般情况下是样本数据，样本数据的协方差为 $\mathrm{cov}(X,Y)=\dfrac{\sum(X-\bar{X})(Y-\bar{Y})}{n-1}$。

用协方差可确定两个变量的变化是否相关（见公式（10-1））。协方差绝对值越大，两个变量的关联越强；协方差为零，则说明两个变量互不相关。如果协方差为正，说明两个变量的共变是同向的；协方差为负，则共变是反向的。因为协方差的度量衡单位与两个变量都有关，当两个变量的度量衡单位不同时，难以用协方差表达“相关”关系。

Excel“协方差”工具计算的是总体协方差，所以为了获得样本协方差，需要将 Excel“协方差”结果乘以 $n/(n-1)$。

对于例 10-2 数据，利用 Excel“协方差”工具获得协方差的步骤如下：

①打开光盘文件“10-3 相关与回归工具.xlsx”，单击的“数据→分析→数据分析→协

方差”，单击“确定”，弹出“协方差”对话框（与图 10-4 类似），获得结果见图 10-5 右下方。

	A	B	C	D	E	F	G
1							
2						X	Y
3	编号	X	Y		X	1	
4	1	13	3.54		Y	0.88177	1
5	2	11	3.01				
6	3	9	3.09				
7	4	6	2.48				
8	5	8	2.56			X	Y
9	6	10	3.36		X	5.25	
10	7	12	3.18		Y	0.730625	0.130773
11	8	7	2.65				

图 10-5 “协方差”工具的分析结果

②获得总体协方差 cov(*X*,*Y*)=0.7306，*X*、*Y* 的总体方差分别为 cov(*X*,*X*)=5.25、cov(*Y*,*Y*)=0.1308。与“相关系数”分析工具一样，也以“下三角阵”的形式给出协方差。由式（10-1）得相关系数：

$$r=\operatorname{cov}(X,Y)/\sqrt{\operatorname{cov}(X,X)\operatorname{cov}(Y,Y)}=0.7306/\sqrt{5.25\times 0.1308}=0.8818$$

由此结果可见，协方差与相关系数间的密切关系。

本例的样本协方差 cov(*X*，*Y*)=0.8350，*X*、*Y* 的样本方差分别为 cov(*X*，*X*)=6.0000、cov(*Y*，*Y*)=0.1495。样本方差的平方根就是通常的“标准差”。

3. 直线回归

利用“回归”数据分析工具对例 10-2 资料进行回归分析的步骤如下：

①打开光盘文件“10-3 相关与回归工具.xlsx”，单击的“数据→分析→数据分析→回归”，单击“确定”，弹出“回归”对话框，如图 10-6 所示，按图 10-6 设置对话框。

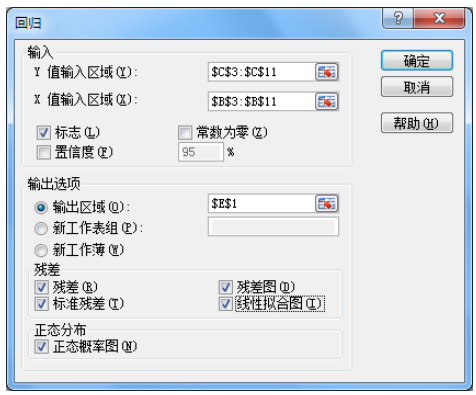


图 10-6 “回归”对话框

②在“回归”对话框中选择“常数为零”，是强制使回归线通过原点（0,0），即使截距

a=0，本例不选。选择“置信度”，可自定义输出截距及回归系数的置信区间，在不选的情况下，系统默认置信度为 95%，本例不选，即只输出 95%的置信区间。数据区域包括变量名称，则选择“标志”。输出区域本例选择在单元 E6 的右下方。

残差是应变量与相应的预测值之差，即 $Y-\hat{Y}$。“标准残差”也叫 Pearson 残差，公式为 $(Y-\hat{Y})/S_{Y-\hat{Y}}$（注意 $S_{Y-\hat{Y}}=S_{Y.X}\left(\frac{n-2}{n-1}\right)$，是残差这列数据的标准差，其中 $S_{Y.X}$ 见公式(10-5)），其均数为 0，标准差为 1。“残差图（residual plot）”是以自变量 X 为横轴，残差为纵轴绘制的散点图。在同一张图上，“线性拟合图（line fit plot）”以自变量 X 为横轴，分别以预测值 $\hat{Y}$ 和观察值 Y 为纵轴绘制的图形。“正态概率图”的横轴为由小到大顺序给出的各个观察值 Y 的百分比排位（可由 Excel 函数 PERCENTRANK(array,x,significance)或=(RANK(x)-1)/(n-1)计算，其中的“x”为被排序的值，“array”为所有数据的区域，“significance”表示返回的百分数值的有效位数），纵轴为观察值 Y，如果 Y 服从正态分布，则正态概率散点图应成一条直线。单击“确定”，可获得如图 10-7 所示的结果。

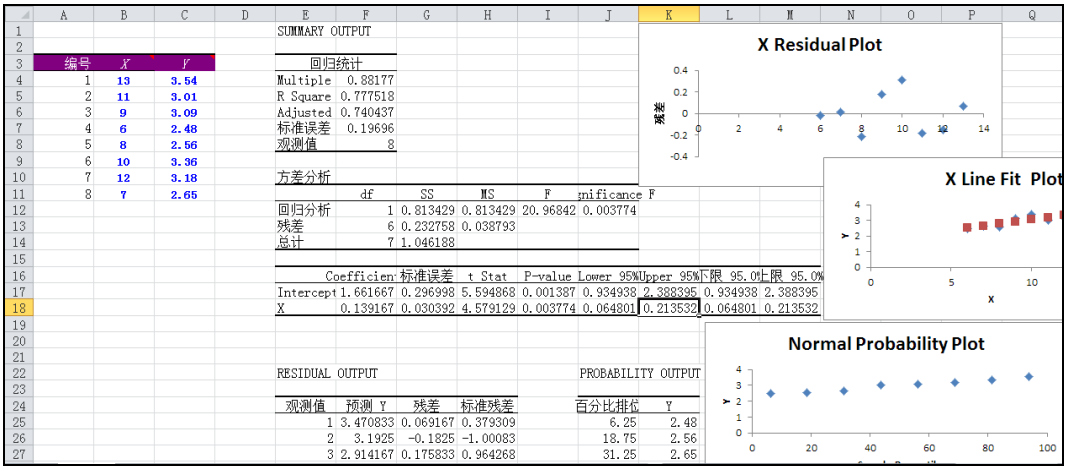


图 10-7　回归分析数据及输出结果

结果分析：由 Excel“回归”分析工具获得的结果，包含很多多重回归的指标。具体阐述如下：

①回归统计量：这一部分结果主要反映模型的拟合优劣程度。复相关系数 R（multiple R），对于简单相关与回归分析，等价于简单 Pearson 相关系数的绝对值（见图 10-7）；决定系数 R^2（R Square）与剩余标准差（标准误差），样本含量（观测值）即为参加计算的 X、Y 对子数，前面已叙述。

校正 R^2（Adjusted R Square）校正了自变量越多，R^2 一定大，方程一定好的弊病。令 p

为自变量的个数，n 为样本含量，则有：

$$校正\ R^2 = 1 - \frac{(1-R^2)(n-1)}{n-p-1} = 1 - \frac{MS_{剩}}{MS_{总}} \qquad (10\text{-}16)$$

$MS_{剩}$ 与 $MS_{总}$ 见方差分析表（表 10-1）。本例有：

$$校正\ R^2 = 1 - \frac{(1-0.8818)(8-1)}{8-1-1} = 0.7775$$

复相关系数、决定系数、校正决定系数越大回归模型越好，剩余标准差越小，模型估计精度越高。

②方差分析：将总变异分解为回归平方和和残差平方和，对整个回归方程进行检验。当只有一个自变量时，其方差分析结果与 t 检验等价，即 $\sqrt{F} = t_b = t_r$。

③回归系数分析：给出了回归方程的系数以及检验结果。由此结果可建立回归方程为：

$$\hat{Y} = 1.6617 + 0.1392X$$

图 10-7 的回归分析结果表中，还给出了回归系数的 95%置信区间（注意：Excel 给出了两组置信区间，原因是一组为默认 95%置信区间，另一组为自定义的置信区间，后者可在图 10-6 中设定）。

④此外也给出了残差输出表（RESIDUAL OUTPUT）、概率输出表（PROBABILITY OUTPUT）、回归图表。其原理已在前面给出。

10.1.3 添加趋势线实现直线回归分析

回归分析还可以用作图的方法实现，即先绘制两个变量的散点图，再通过添加趋势线来实现回归分析。这种方式简单快捷，不仅适合于直线回归分析，同样适合于下一节将要介绍的非线性回归分析。Excel 具体实施步骤如下：

①输入原始数据，作 *XY* 散点图（见光盘文件“10-4 回归趋势线.xlsx”）。

②在散点图的任一个点上单击鼠标右键，选择快捷菜单中的“添加趋势线”，弹出的“设置趋势线格式”对话框如图 10-8 所示。

③单击左侧“趋势线选项”标签，在“趋势预测/回归分析类型”下选择“线性”。

④选中“显示公式”和“显示 R 平方值”复选框（见图 10-8 的下方）。为了使在图中绘出的趋势线沿着 X 轴向前或/和向后延伸，可在“前推”或/和“倒推”框中键入数字；如果要求回归线过某一点（0,a），可在“设置截距(S)=”后的文本框中键入“a”值。

⑤单击“关闭”，所得结果如图 10-9 所示。

如果要显示更多小数位数，可右键单击选取散点图中含有方程和 R^2 的区域，选择“设置趋势线标签格式”，单击左侧“数字”标签，对小数位进行调整。对于这一区域中的文本可用通常的文本编辑方法进行编辑。

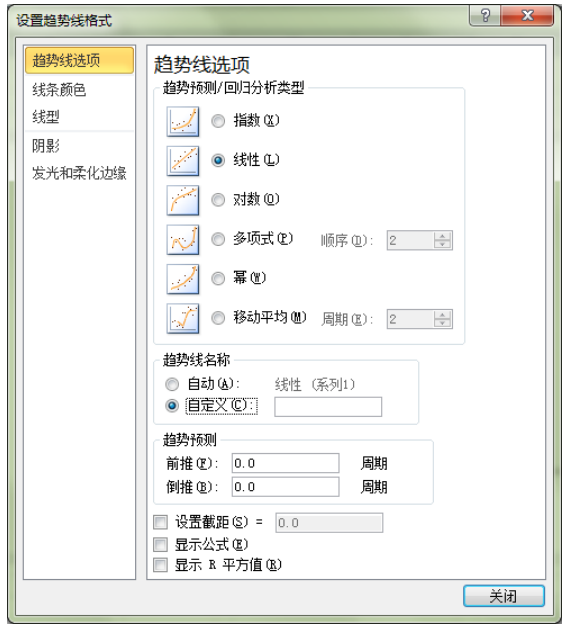


图 10-8 “添加趋势线”对话框

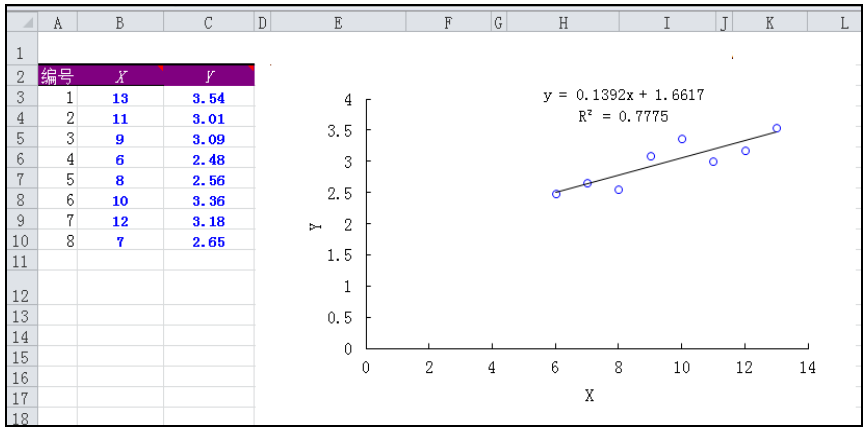


图 10-9 添加趋势线后的结果

10.2 曲线回归

有时，两个变量之间并非一定呈线性关系，而是呈现出某种曲线趋势，此时需要拟合曲线方程，拟合这种曲线的方法称为曲线回归（curve regression）。常见的曲线函数形式有：

①对数函数形式：$Y = a + b \cdot \ln(X)$

②乘幂函数形式：$Y = aX^b$ 或 $\ln(Y) = \ln(a) + b\ln(X)$

③指数函数形式：$Y = a \cdot e^{bX}$ 或 $\ln(Y) = \ln(a) + bX$

④多项式形式：$Y = a + b_1X + b_2X^2 + \cdots + b_nX^n$

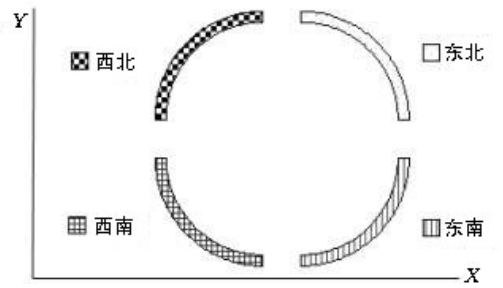


图 10-10　非线性曲线样式图示

曲线拟合时应根据 X 与 Y 的散点分布形状选择函数类型。当散点图显示曲线凸面朝向“西北”方向（见图 10-10）时，可选用乘幂函数（$X > 1$）、对数函数；当曲线凸面朝向“西南”方向时，可选用乘幂函数、对数函数或指数函数；当曲线凸面朝向“东南”方向时，可选用乘幂函数（$X > 1$）或指数函数。此外，这 4 种图形都可以选用二次多项式函数。如果散点图显示曲线有两个凸面（S 形），那么可选用三次多项式函数。

以上表明，一种曲线样式可以采用多种曲线函数形式，究竟应选取哪一种函数模型，可比较各曲线的决定系数 R^2 作出决定。决定系数 R^2 越大，模型越好。

与直线回归一样，也可以用两种方法进行曲线回归拟合，即添加趋势线法与“回归”分析工具法。后者需要通过变量变换，使变换后的变量间呈直线关系，再利用“回归”分析工具进行拟合；为了获得曲线方程，还需采用反变换的方法得到关于原变量 Y 和 X 的曲线方程。

10.2.1　对数曲线拟合

例 10-4：用已知浓度 X 的免疫球蛋白 A(IgA, μg/ml)作火箭电泳，测得火箭高度 Y(mm)如下。试拟合 Y 关于 X 的非线性回归方程。

IgA（μg/ml）X	0.2	0.4	0.6	0.8	1.0	1.2	1.4	1.6
火箭高度（mm）Y	7.6	12.3	15.7	18.2	18.7	21.4	22.6	23.8

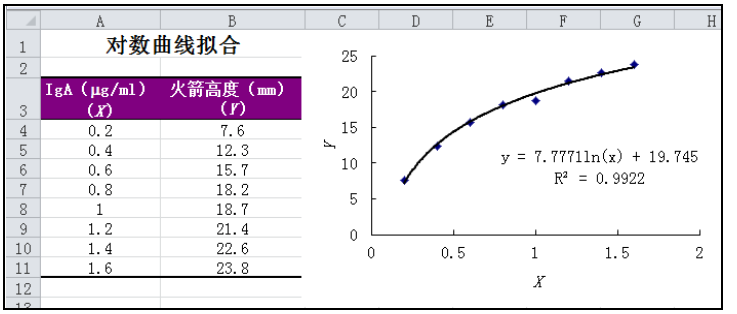
1．添加趋势线

添加趋势线拟合对数曲线模型的简略步骤如下：

①按图 10-11（也见光盘文件“10-5 对数曲线回归.xlsx”）键入数据，并绘制散点图。

②在散点图的任一个散点上单击鼠标右键，选择快捷菜单中的“添加趋势线”，在弹出的“设置趋势线格式”对话框单击左侧“趋势线选项”标签，选择“对数”类型。

③选中“显示公式”和“显示 R 平方值”复选框，单击“关闭”。由此获得对数曲线回归方程为 $\hat{Y}=19.745+7.7771\ln(X)$，$R^2$=0.9922，如图 10-11 所示。



	A	B
1	对数曲线拟合	
2		
3	IgA（μg/ml）（X）	火箭高度（mm）（Y）
4	0.2	7.6
5	0.4	12.3
6	0.6	15.7
7	0.8	18.2
8	1	18.7
9	1.2	21.4
10	1.4	22.6
11	1.6	23.8
12		

图 10-11　添加对数趋势线结果

结果分析：如果分别拟合线性、乘幂、指数曲线，将获得决定系数 R^2 分别为 0.9391、0.9823、0.8425，都小于对数曲线回归的决定系数 R^2=0.9922，因此该资料拟合对数曲线较好。

2. 回归工具

采用“回归”分析工具拟合对数曲线模型的步骤如下：

①在单元格 C4 中输入公式“=LN(A4)”，将原始数据的自变量 X 作对数变换，即计算 $\ln(X)$。将此公式向下拖放填充至单元格 C11。

②单击“数据→分析→数据分析→回归”，在“回归”对话框中的“Y 值输入区域”选择“B4:B11”，“X 值输入区域”选择“C4: C11”，即用第 C 列和 B 列数据分别当作 X 和 Y 进行线性回归分析。“输出区域”选取单元格 E1，获得的回归分析结果如图 10-12 所示。

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	对数曲线拟合				SUMMARY OUTPUT								
2													
3	IgA（μg/ml）（X）	火箭高度（mm）（Y）	ln(X)		回归统计								
4	0.2	7.6	-1.60944		Multiple	0.996094							
5	0.4	12.3	-0.91629		R Square	0.992203							
6	0.6	15.7	-0.51083		Adjusted	0.990903							
7	0.8	18.2	-0.22314		标准误差	0.523793							
8	1	18.7	0		观测值	8							
9	1.2	21.4	0.182322										
10	1.4	22.6	0.336472		方差分析								
11	1.6	23.8	0.470004			df	SS	MS	F	gnificance F			
12					回归分析	1	209.4726	209.4726	763.4989	1.48E-07			
13					残差	6	1.646152	0.274359					
14					总计	7	211.1188						
15													
16						Coefficien	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
17					Intercept	19.74512	0.201688	97.89932	7.65E-11	19.251603	20.2386288	19.2516	20.23863
18					ln(X)	7.77706	0.281456	27.63148	1.49E-07	7.0883608	8.46575915	7.088361	8.465759

图 10-12　“回归”工具获得的对数曲线模型拟合结果

获得的回归系数及决定系数结果与“添加趋势线”一致。此外，还获得了回归系数检验结果、剩余标准差等。

10.2.2 乘幂曲线拟合

例 10-5：某地大气中氟化物测定结果如下，试拟合一条曲线。

染源距离 X	50	100	150	200	250	300	400	500
氟化物浓度 Y	0.687	0.398	0.200	0.121	0.090	0.050	0.020	0.010

1. 添加趋势线

添加趋势线拟合乘幂曲线模型的简略步骤如下：

①按图 10-13（也见光盘文件“10-6 乘幂曲线回归.xlsx”）键入数据，并绘制散点图。

②在散点图的任一个散点上单击鼠标右键，选择快捷菜单中的“添加趋势线”，在弹出的“设置趋势线格式”对话框中单击左侧“趋势线选项”标签，选择“乘幂”类型。

③选中“显示公式”和“显示 R 平方值”复选框，单击“关闭”。由此获得乘幂曲线回归方程为 $\hat{Y}=1509X^{-1.8338}$，R^2=0.9324，如图 10-14 所示。

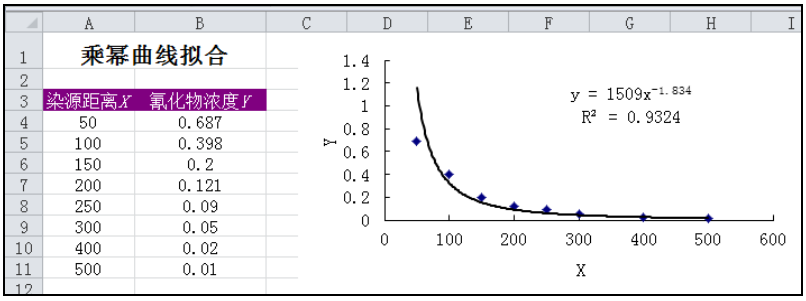


图 10-13 添加乘幂趋势线结果

结果分析：如果分别拟合线性、对数、指数曲线，将获得决定系数 R^2 分别为 0.6739、0.9248、0.9915，与乘幂曲线回归的决定系数 R^2=0.9324 比较，该资料拟合指数曲线更合适。

2. 回归工具

采用“回归”分析工具拟合对数曲线模型的步骤如下：

①在单元格 C4 中输入公式“=LN(A4)”，将原始数据的自变量 X 作对数变换，即计算 ln(X)。将此公式向下拖放填充至单元格 C11。

②在单元格 D4 中输入公式“=LN(B4)”，将原始数据的应变量 Y 作对数变换，即计算 ln(Y)。将此公式向下拖放填充至单元格 D11。

③单击的“数据→分析→数据分析→回归”，在“回归”对话框中的“Y 值输入区域”

选择“D4:D11”，“X 值输入区域”选择“C4:C11”，即用第 C 列和 D 列数据分别当作 X 和 Y 进行线性回归分析。“输出区域”选择单元格 F1，获得的回归分析结果如图 10-14 所示。

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	乘幂曲线拟合					SUMMARY OUTPUT								
2														
3	染源距离X	氟化物浓度Y	ln(X)	ln(Y)		回归统计								
4	50	0.687	3.912023	-0.37542		Multiple	0.965608							
5	100	0.398	4.60517	-0.9213		R Square	0.932398							
6	150	0.2	5.010635	-1.60944		Adjusted	0.921131							
7	200	0.121	5.298317	-2.11196		标准误差	0.404037							
8	250	0.09	5.521461	-2.40795		观测值	8							
9	300	0.05	5.703782	-2.99573										
10	400	0.02	5.991465	-3.91202		方差分析								
11	500	0.01	6.214608	-4.60517			df	SS	MS	F	gnificance F			
12						回归分析	1	13.50937	13.50937	82.75458	9.91E-05			
13						残差	6	0.979477	0.163246					
14						总计	7	14.48885						
15														
16							Coefficien	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
17						Intercept	7.319203	1.074355	6.812651	0.00049	4.6903521	9.94805402	4.690352	9.948054
18						ln(X)	-1.83382	0.201586	-9.09695	9.91E-05	-2.327085	-1.3405572	-2.32708	-1.34056

图 10-14 “回归”工具获得的乘幂曲线模型拟合结果

拟合出的线性方程为 $\ln(\hat{Y})=\ln(a)+b\ln(X)=7.3192-1.8338\ln(X)$，变换为幂函数方程为 $\hat{Y}=\mathrm{e}^{7.3192}\cdot X^{-1.8338}=1509X^{-1.8338}$，决定系数 R^2=0.9324。这些结果与添加趋势线一致。此外，还获得回归系数的 t 检验结果 P=9.9097×10^{-5}<0.05，说明回归方程有统计学意义，该回归模型的剩余标准差为 0.4040。

10.2.3 指数曲线拟合

下面利用例 10-5 数据拟合指数曲线，步骤同前，只是在添加趋势线时选择趋势线的类型为“指数”，得到的结果见图 10-15。拟合出的指数方程为 $\hat{Y}=0.9293\mathrm{e}^{-0.0094X}$，决定系数 R^2=0.9915。

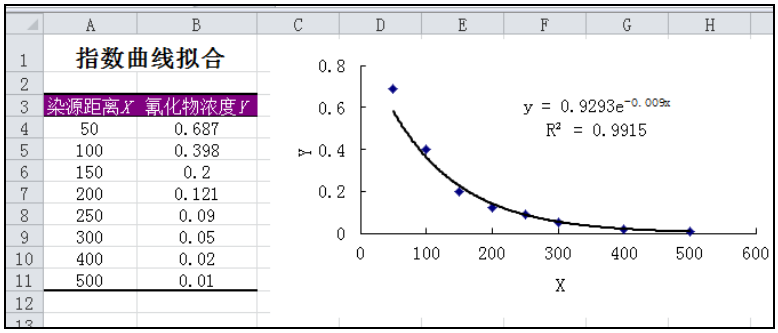


图 10-15 添加指数趋势线结果

再用“回归”分析工具进行分析。在单元格 C4 中键入公式“=LN(B4)”，并拖放填充至 C11，计算应变量 Y 的对数 ln(Y)。以单元格区域 A4:A11 中的自变量 X 数据与单元格区

域 C4:C11 中的 ln(*Y*)数据拟合线性回归模型，得到回归分析结果如图 10-16（也见光盘文件“10-7 指数曲线回归.xlsx”）所示。

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	指数曲线拟合				SUMMARY OUTPUT								
2													
3	染源距离*X*	氧化物浓度*Y*	Ln(Y)		回归统计								
4	50	0.687	-0.3754		Multiple	0.995752							
5	100	0.398	-0.9213		R Square	0.991522							
6	150	0.2	-1.6094		Adjusted	0.990109							
7	200	0.121	-2.112		标准误差	0.143085							
8	250	0.09	-2.4079		观测值	8							
9	300	0.05	-2.9957										
10	400	0.02	-3.912		方差分析								
11	500	0.01	-4.6052			df	SS	MS	F	gnificance F			
12					回归分析	1	14.36601	14.36601	701.6951	1.9105E-07			
13					残差	6	0.12284	0.020473					
14					总计	7	14.48885						
15													
16						Coefficien	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
17					Intercept	-0.07332	0.100295	-0.73103	0.492311	-0.3187322	0.17209494	-0.31873	0.172095
18					染源距离X	-0.00941	0.000355	-26.4895	1.91E-07	-0.0102809	-0.0085421	-0.01028	-0.00854

图 10-16 “回归”工具获得的指数曲线模型拟合结果

拟合出的线性方程为 $\ln(Y)=-0.0733-0.0094X$，变换后的指数曲线方程为 $\hat{Y}=\exp(-0.0733-0.0094X)=0.9293e^{-0.0094X}$，决定系数 $R^2=0.9915$。结果与添加趋势线一致。此外，回归系数的 *t* 检验结果显示 $P=1.9106\times10^{-7}<0.05$，说明回归模型有统计学意义。该模型的剩余标准差为 0.1431。

与乘幂曲线模型相比，拟合指数曲线的决定系数 R^2 更大，回归的剩余标准差更小，说明该资料拟合指数曲线的效果更好。

10.2.4 多项式曲线拟合

例 10-6：在不同缺氧程度（mmHg）下，大白鼠受相同剂量放射线照射后，骨髓内坏死灶（个/50 视野）亦不同，资料如下。试拟合一条回归曲线。

缺氧程度（mmHg）*X*	300	400	500	600	700	760
坏死灶（个/50 视野）*Y*	214.5	87.5	92.5	136.6	180.2	212.0

1. 添加趋势线

添加趋势线拟合多项式曲线模型的简略步骤如下：

① 按图 10-17（也见光盘文件“10-8 多项式回归.xlsx”）键入数据，并绘制散点图。

② 在散点图的任一个散点上单击鼠标右键，选择快捷菜单中的“添加趋势线”，在弹出的“设置趋势线类型”类型对话框中单击左侧“趋势线选项”标签，选择“多项式”类型。此时旁边的“顺序”下拉菜单激活，可设置多个顺序（最大为 6），本例选定 2。

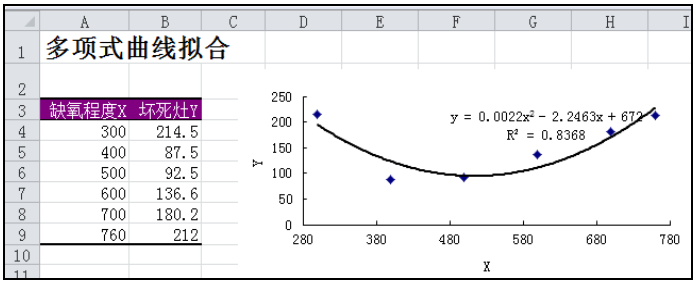


图 10-17　添加多项式趋势线结果

③选中“显示公式”和“显示 R 平方值”复选框，单击“关闭”后，获得多项式曲线回归方程为 $\hat{Y} = 672 - 2.2463X + 0.0022X^2$，$R^2$=0.8368，如图 10-17 所示。

选中趋势线，按 Delete 键，可同时删除趋势线及其方程与决定系数。这种情况下可反复考虑拟合其他类型的曲线，本例拟合线性、对数、乘幂、指数、三次多项式、四次多项式后，分别获得决定系数 R^2 为 0.0683、0.019、0.0361、0.0953、0.9851、1。尽管三次多项式、四次多项式的 R^2 很大，但增加了模型应用的难度，从整体上考虑，下面仍针对二次多项式，采用“回归”数据分析工具拟合一般线性模型。

2. 回归工具

采用“回归”分析工具拟合多项式曲线模型的步骤如下：

①将应变量 Y 放在第 A 列，自变量 X 放在第 B 列，在单元格 C4 中输入公式“=B4^2”，向下拖放填充至 C9，计算自变量 X 的平方 X^2。

②单击“数据→分析→数据分析→回归”，在“回归”对话框的“Y 值输入区域”中选择“A4:A9”，“X 值输入区域”选择“B4:C9”，即用第 B 列和 C 列数据(X, X^2)分别当作自变量与第 A 列的应变量 Y 进行线性回归分析。“输出区域”选择单元格 E1，获得的回归分析结果如图 10-18 所示。这些结果与添加趋势线获得结果一致。

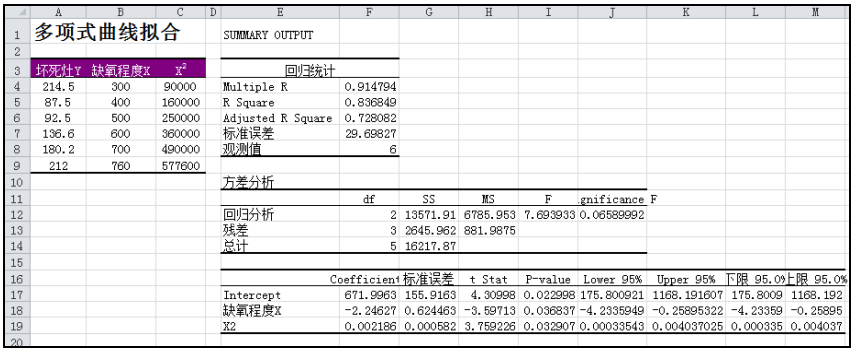


图 10-18　“回归”工具获得的多项式拟合结果

10.3 非线性回归

以上曲线回归仍然采用了最小二乘法原理，也就是使残差平方和$\sum\left(Y-\hat{Y}\right)^2$最小来获得回归参数估计值$a$、$b$。对于乘幂曲线和指数曲线，对应变量$Y$进行了对数变换，这两个曲线只能使$\sum\left(\ln(Y)-\ln(\hat{Y})\right)^2$最小，并不能保证使$\sum\left(Y-\hat{Y}\right)^2$最小。所以，凡对应变量$Y$进行变换的曲线回归结果只能作为应用的参考，而不能作为最终结果，最终方程应采用本节结果。下面仍利用10.2节所给的例子，采用工具菜单中的“规划求解参数”进行参数求解。

10.3.1 对数曲线拟合

规划求解对数曲线方程的具体步骤如下：

①按图10-19（也见光盘文件“10-9 对数曲线.xlsx”）在A4:B2003范围内键入数据。

	A	B	C	D	E	F	G
1			规划求解对数曲线方程		$\hat{Y}=a+b\cdot\ln(X)$		
2							
3	X	Y	Y^	残差平方(Y-Y^)²	残差平方和	a	b
4	0.2	7.6	7.228417505	0.13807355	1.64615241	19.745113	7.77706
5	0.4	12.3	12.61906489	0.101802404			
6	0.6	15.7	15.77239146	0.005240524			
7	0.8	18.2	18.00971227	0.036209419			
8	1	18.7	19.74511312	1.092261427			
9	1.2	21.4	21.16303885	0.056150587			
10	1.4	22.6	22.36187797	0.0567021			
11	1.6	23.8	23.40035966	0.159712402			

图10-19 规划求解对数曲线

②在单元格C4中键入公式“=IF(AND(A4<>"",B4<>""),$F$4+$G$4*LN(A4),"")”，并将此公式拖放至C2003，由公式$\hat{Y}=a+b\cdot\ln(X)$计算应变量Y的预测值$\hat{Y}$；在单元格D4中键入公式“=IF(AND(A4<>"",B4<>""),(B4-C4)^2,"")”，并将此公式拖放至D2003，由公式$(Y-\hat{Y})^2$计算残差的平方；在单元格E4中键入公式“=SUM(D:D)”计算残差平方和。

③单击“数据→分析→规划求解”，在弹出的“规划求解参数”对话框（见图10-20）中，设置目标“E4”，并使该单元格的值为“最小值”，可变单元格为“F4:G4”。单击“求解”，便可使残差平方和最小（即最小二乘法原理）求解回归参数a（单元格F4）、b（单元格G4）（见图10-19）。

结果分析：规划求解获得a=19.7451、b=7.7771，残差平方和（单元格E4）为1.6462，与10.2.1节的图10-13中结果相同。原因是对数曲线只对X作变换而未对Y作变换，所以可以保证使$\sum\left(Y-\hat{Y}\right)^2$最小。

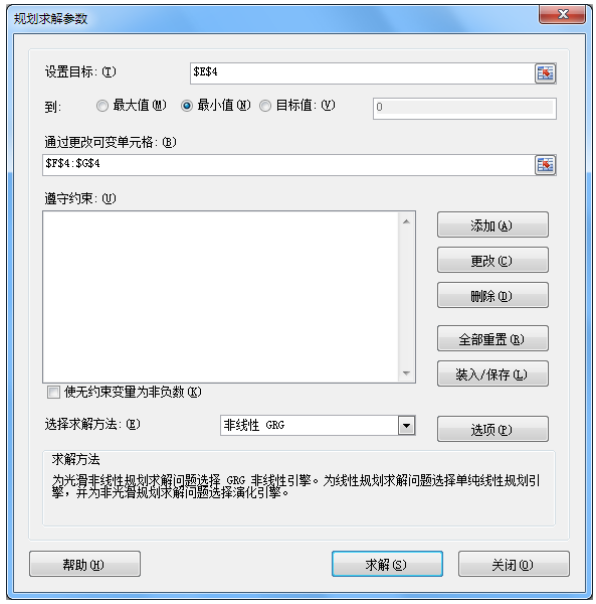


图 10-20 “规划求解参数”对话框

温馨提示：如果需要进行对数曲线拟合，用户可将数据输入到 B4:C2003 范围内，然后清除单元格 F4、G4 中的数字，按图 10-20 调用“规划求解”便可完成参数求解。

10.3.2 乘幂曲线拟合

规划求解乘幂曲线方程的步骤与对数曲线基本相同，不同的是：在单元格 C4 中键入公式“=IF(AND(A4<>"",B4<>""),$F$4*A4^$G$4,"")”，并将此公式拖放至 C2003，由公式 $\hat{Y}=aX^b$ 计算应变量 Y 的预测值 $\hat{Y}$。规划求解获得的结果见图 10-21（也见光盘文件“10-10 乘幂曲线.xlsx”）。

	A	B	C	D	E	F	G
1			规划求解乘幂曲线方程		$\hat{Y}=aX^b$		
2							
3	X	Y	Y^	残差平方(Y-Y^)²	残差平方和	a	b
4	50	0.687	0.711035815	0.00057772	0.0126061	73.640434	-1.18615
5	100	0.398	0.312482886	0.007313177			
6	150	0.2	0.193177464	4.6547E-05			
7	200	0.121	0.137328601	0.000266623			
8	250	0.09	0.105392975	0.000236944			
9	300	0.05	0.084896781	0.001217785			
10	400	0.02	0.060352568	0.00162833			
11	500	0.01	0.04631764	0.001318971			

图 10-21 规划求解乘幂曲线

结果分析：规划求解获得 a=73.6404、b=–1.1862（与图 10-14 不同）。残差平方和（单元格 E4）为 0.0126，与 10.2.2 节的图 10-14 中结果方程 $\hat{Y}=e^{7.3192}\cdot X^{-1.8338}=1509X^{-1.8338}$ 计算出的残差平方和=1.8721 相比要小。原因是乘幂曲线对 Y 作了变换，所以不能保证使 $\sum\left(Y-\hat{Y}\right)^2$ 最小。

温馨提示：如果需要进行乘幂曲线拟合，用户可将数据输入到 B4:C2003 范围内，然后清除单元格 F4、G4 中的数字，按图 10-20 调用“规划求解”便可完成参数求解。

10.3.3 指数曲线拟合

规划求解指数曲线方程的步骤与对数曲线基本相同，不同的是：在单元格 C4 中键入公式“=IF(AND(A4<>"",B4<>""),$F$4*EXP($G$4*A4),"")”，并将此公式拖放至 C2003，由公式 $\hat{Y}=a\cdot e^{bX}$ 计算应变量 Y 的预测值 $\hat{Y}$。规划求解获得的结果见图 10-22（也见光盘文件“10-11 指数曲线.xlsx”）。

	A	B	C	D	E	F	G
1			规划求解指数曲线方程	$\hat{Y}=a\cdot e^{bX}$			
2							
3	X	Y	Y^	残差平方(Y-Y^)²	残差平方和	a	b
4	50	0.687	0.686517384	2.32918E-07	0.00106339	1.2089094	-0.01132
5	100	0.398	0.389860583	6.62501E-05			
6	150	0.2	0.221394648	0.000457731			
7	200	0.121	0.125725944	2.23345E-05			
8	250	0.09	0.071397449	0.000346055			
9	300	0.05	0.040545297	8.93914E-05			
10	400	0.02	0.013075432	4.79496E-05			
11	500	0.01	0.00421669	3.34467E-05			

图 10-22 规划求解指数曲线

结果分析：规划求解获得 $a=1.2089$、$b=-0.0113$（与图 10-16 不同）。残差平方和（单元格 E4）为 0.0011，与 10.2.3 节的图 10-16 中结果方程 $\hat{Y}=\exp(-0.0733-0.0094X)=0.9293e^{-0.0094X}$ 计算出的残差平方和=0.0138 相比要小。原因是指数曲线对 Y 作了变换，所以不能保证使 $\sum\left(Y-\hat{Y}\right)^2$ 最小。

温馨提示：如果需要进行指数曲线拟合，用户可将数据输入到 B4:C2003 范围内，然后清除单元格 F4、G4 中的数字，按图 10-20 调用“规划求解”便可完成参数求解。

10.3.4 多项式曲线拟合

规划求解二次多项式曲线方程的步骤与对数曲线基本相同，不同的是：在单元格 C4

中键入公式“=IF(AND(A4<>"",B4<>""),$F$4+$G$4*A4+$H$4*A4^2,"")”，并将此公式拖放至 C2003，由公式 $\hat{Y}=a+b_1X+b_2X^2$ 计算应变量 Y 的预测值 $\hat{Y}$。规划求解获得的结果见图 10-23（也见光盘文件“10-12 多项式.xlsx”）。

	A	B	C	D	E	F	G	H
1	规划求解二次多项式方程 $\hat{Y}=a+b_1X+b_2X^2$							
2								
3	X	Y	Y^	残差平方(Y-Y^)²	残差平方和	a	b_1	b_2
4	300	214.5	194.4435483	402.2612554	2646.52117	668.18362	-2.23061	0.002172
5	400	87.5	123.3943006	1288.400816				
6	500	92.5	95.77710702	10.7394304				
7	600	136.6	111.5919675	625.4016892				
8	700	180.2	170.8388821	87.63052869				
9	760	212	227.2344168	232.0874548				

图 10-23　规划求解多项式曲线

结果分析：规划求解获得 a=668.1836、b_1=−2.2306、b_2=0.0021（与图 10-17、图 10-18 十分相似，应该完全相同，可能算法不同导致个数或小数位数值略有差异）。残差平方和（单元格 E4）为 2646.5212，也与 10.2.4 节的图 10-18 方差分析表的 $SS_{残差}$（单元格 G13）相同。原因是多项式并不对 Y 作变换，所以能保证使 $\sum\left(Y-\hat{Y}\right)^2$ 最小。

温馨提示：如果需要进行多项式曲线拟合，用户可将数据输入到 B4:C2003 范围内，然后清除单元格 F4、G4 中的数字，按图 10-20 调用“规划求解”便可完成参数求解。

由以上分析可见，只有对应变量 Y 作变换的曲线才需要采用本节结果，如果不对应变量 Y 作变换，则应采用 10.2 节的结果。

10.4　多重相关与回归

许多情况下，需要考虑两个以上的观测变量，此时可用多重相关（multiple correlation）与多重回归（multiple regression）来分别定量反映一个变量与一组自变量 X_1、X_2 、…、X_p 之间的线性相关关系与线性依存关系。

10.4.1　多重相关与多重协方差

多重相关要求 X_1、X_2、…、Y 为多变量正态分布（multivariate normaldistribution）的随机变量。通常用复相关系数（multiple correlation coefficient，Excel 称为 Multiple R）来度量某个随机变量 Y 和 p 个随机变量（X_1、X_2 、…、X_p）之间线性联系的程度。

多重回归只要求 Y 为正态分布的随机变量。Y 关于 X_1、X_2 、…、X_p 的多重回归方程可写为：

$$\hat{Y}=b_0+b_1X_1+b_2X_2+\cdots+b_pX_p \qquad (10\text{-}17)$$

应变量 Y 与估计值 $\hat{Y}$ 之间的简单相关系数等于复相关系数 R，其取值范围为 $0<R<1$。因为 $\hat{Y}$ 包括了所有自变量 X 的信息。关于复相关系数 R 是否为零的假设检验，等价于多重回归中检验整个回归方程的方差分析，参见 10.4.2 节“多重回归分析”。

对于多个变量，两两间的相关系数（常称为简单相关系数）可采用“相关系数”数据分析工具获得。

偏相关系数可用来描述扣除其他变量影响之后，应变量与其他自变量之间的线性相关关系。本书不作介绍，具体计算参见有关统计书籍。

对于多个变量，两两间的协方差可采用“协方差”数据分析工具获得。

例 10-7：第 6 章的例 6-8 给出了随机挑选的 18 名儿童的身高、体重、体表面积测量值，试分析他们之间的相关性，并建立体表面积依赖于身高、体重的线性回归方程。

解：多重相关与多重协方差分析可简单按如下步骤进行。

① 按图 10-24（见光盘文件“10-13 多重相关与多重协方差.xlsx”）在第 A、B、C 列分别键入体重、身高、体表面积等 18 行数据。

	A	B	C	D	E	F	G	H	I
1	多重相关与多重协方差								
2									
3	体重(kg)X_1	身高(cm)X_2	表面积(m^2)Y						
4	3	51	0.1850						
5	8	77	0.3934			体重(kg)X1	身高(cm)X2	表面积(m2)Y	
6	15	94	0.6075		体重(kg)X1	1			
7	16	105	0.6785		身高(cm)X2	0.97133628	1		
8	20	114	0.7433		表面积(m2)Y	0.99265856	0.99137657	1	
9	25	130	0.9028						
10	32	140	1.0584						
11	32	144	1.0822			体重(kg)X1	身高(cm)X2	表面积(m2)Y	
12	41	153	1.2627		体重(kg)X1	146.521798			
13	2	51	0.1928		身高(cm)X2	377.01929	1028.21914		
14	10	76	0.3846		表面积(m2)Y	4.02700904	10.6540096	0.11232158	
15	16	99	0.6411						
16	16	105	0.6407						
17	25	123	0.8756			155.140727			
18	25	124	0.8855			399.196895	1088.70261		
19	30	141	1.0291			4.26389192	11.280716	0.11892873	
20	36	143	1.1282						
21	45	157	1.3427						

图 10-24　多重相关与多重协方差分析结果

② 单击“数据→分析→数据分析→相关系数”，单击“确定”，弹出“相关系数”对话框，设置“输入区域”为“A3:C21”，“分组方式”为逐列，选择标志位于第一行，“输出区域”为 E5。单击“确定”，获得身高与体重、表面积与体重、表面积与身高两两之间的 Pearson 相关系数分别为 0.9713、0.9927、0.9914。

③ 单击“数据→分析→数据分析→协方差”，单击“确定”，弹出“协方差”对话框，设置“输入区域”为“A3:C21”，“分组方式”为逐列，选择标志位于第一行，“输出区域”为 E11。单击“确定”，获得身高与体重、表面积与体重、表面积与身高两两之间的总体协方差分别为 377.0193、4.0270、10.6540；体重、身高、体表面积的总体方差分别为 146.5218、

1028.2191、0.1123。为了获得样本方差与协方差，需要将由“协方差”工具得到的数值乘以 $n/(n-1)$。本例应乘以 18/17，得到身高与体重、表面积与体重、表面积与身高两两之间的样本协方差分别为 399.1969、4.2639、11.2807；体重、身高、体表面积的样本方差分别为 155.1407、1088.7026、0.1189（参见 10.1.2 节的第 2 点协方差）。

结果分析：本例体重、身高、表面积两两之间的相关系数与协方差都为正，且相关系数绝对值较大，说明两两之间有较高的正相关关系。

10.4.2 多重回归分析

10.4.1 节给出的公式（10-17）中，b_0 为回归方程常数项，又称截距；b_0、b_1、…、b_P 称为偏回归系数（partial regression coefficient），简称回归系数，b_i 表示在除 X_i 以外的其他自变量固定的条件下，X_i 每改变一个单位后 Y 的平均变化量。

评价多重回归模型拟合效果也可用决定系数 R^2、校正决定系数（见公式（10-16））和剩余标准差。通过对回归方程的方差分析，可从整体上反映模型有无统计学意义。

对例 10.7 中数据进行多重回归分析可按如下步骤进行。

①按图 10-25（见光盘文件“10-14 多重回归.xlsx”）在第 A、B、C 列分别键入体重、身高、体表面积等 18 行数据。

②单击“数据→分析→数据分析→回归”，在“回归”对话框中的“Y 值输入区域”选择“C3:C21，“X 值输入区域”选择“A3:B21”，选择“标志”复选框。即用第 A 列和 B 列数据（体重、身高）分别当作自变量与第 C 列的应变量 Y 进行线性回归分析。“输出区域”选取单元格 E3，获得的回归分析结果如图 10-25 所示。

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	多重回归分析												
2													
3	体重(kg)X_1	身高(cm)X_2	表面积(m^2)Y		SUMMARY OUTPUT								
4	3	51	0.1850										
5	8	77	0.3934		回归统计								
6	15	94	0.6075		Multiple R	0.999218							
7	16	105	0.6785		R Square	0.9984366							
8	20	114	0.7433		Adjusted R Sq	0.9982281							
9	25	130	0.9028		标准误差	0.0145166							
10	32	140	1.0584		观测值	18							
11	32	144	1.0822										
12	41	153	1.2627		方差分析								
13	2	51	0.1928			df	SS	MS	F	gnificance F			
14	10	76	0.3846		回归分析	2	2.0186274	1.0093137	4789.587	9.029E-22			
15	16	99	0.6411		残差	15	0.003161	0.00021073					
16	16	105	0.6407		总计	17	2.0217884						
17	25	123	0.8756										
18	25	124	0.8855			Coefficient	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
19	30	141	1.0291		Intercept	-0.104744	0.0261992	-3.9979953	0.001164	-0.160586	-0.0489	-0.16059	-0.0489
20	36	143	1.1282		体重(kg)X1	0.014552	0.0011891	12.2374833	3.31E-09	0.0120174	0.017087	0.012017	0.017087
21	45	157	1.3427		身高(cm)X2	0.0050258	0.0004489	11.1961184	1.11E-08	0.004069	0.005983	0.004069	0.005983

图 10-25 回归分析数据及输出结果

结果分析：本例获得复相关系数为 0.9992，决定系数 R^2=0.9984，调整决定系数为

0.9982，剩余标准差为 0.0145，模型拟合效果较好。经方差分析得到 F 值为 4789.587，P 值为 9.029×10^{-22}，模型具有统计学意义。由回归系数可建立线性回归方程为：

$$\hat{体表面积} = -0.104744 + 0.014552体重 + 0.0050258身高$$

对体重、身高回归系数检验获得的 P 值分别为 3.30551E-09、1.10844E-08，结论与方差分析结果相似。

为了确定哪个变量对模型的贡献更大，需要计算标准化回归系数。如本例，先分别求出体重、身高、体表面积的标准差，再利用标准化回归系数计算公式 $b'_j = b_j(s_j / s_y)$ 获得标准化回归系数。例如可选取单元格 H1，在其中键入公式“=STDEV(A:A)”，计算体重的标准差（12.4555），将此公式向右拖放填充至 J1，分别获得身高与体表面积的标准差为 32.9954 与 0.3449；分别在单元格 H2 与 I2 中输入“=F20*H1/$J$1”与“=F21*I1/$J$1”，获得体重与身高的标准化回归系数分别为 0.5256、0.4809。因为体重的标准化回归系数较大，说明体重对该线性回归模型的贡献较大。

10.4.3 用函数实现多重回归与相关

光盘文件“10-15 多重回归函数.xlsx”利用 5 个工作表（数据、回归、残差、相关、偏 F 值），进行多重回归分析与多重相关分析，并能获得残差分析图以及偏 F 检验结果。

下面以例 10-7 为例，说明如何使用该程序。

①如图 10-26（也见光盘文件“10-15 多重回归函数.xlsx”）所示为“数据”工作表，用于输入数据。第 4 行用于给出变量名称，如单元格 B4、D4、E4 分别命名应变量 Y、自变量 $X1$、自变量 $X2$ 为表面积、体重、身高。单元格区域 B5:B204 可输入应变量数据，单元格区域 D5:M204 可输入自变量数据。程序设定最大处理自变量个数为 10、最大观察个体为 200。单元格 F1 供输入数据标题用。



多重回归与相关

表面积与体重身高的关系

编号	Y 表面积(m^2)	X1 体重(kg)	X2 身高(cm)	X3	X4	X5	X6	X7	X8	X9	X10
1	0.1850	3	51								
2	0.3934	8	77								
3	0.6075	15	94								
4	0.6785	16	105								
5	0.7433	20	114								
6	0.9028	25	130								
7	1.0584	32	140								
8	1.0822	32	144								
9	1.2627	41	153								
10	0.1928	2	51								
11	0.3846	10	76								
12	0.6411	16	99								
13	0.6407	16	105								
14	0.8756	25	123								
15	0.8855	25	124								
16	1.0291	30	141								
17	1.1282	36	143								
18	1.3427	45	157								

图 10-26 “数据”工作表

②在“回归”工作表（见图 10-27）中回归系数及其 t、F 检验结果与图 10-25 结果完

全相同。此外，该程序在给定所有自变量 X 值条件下，估计了 Y 个体值与均数的 $100(1-\alpha)\%$ 置信区间。注意单元格区域 C22:L22 为用户给定 X 值的区域。置信度 $100(1-\alpha)\%$ 也可人为修订。实际意义与前面相同，所以此处不再赘述。

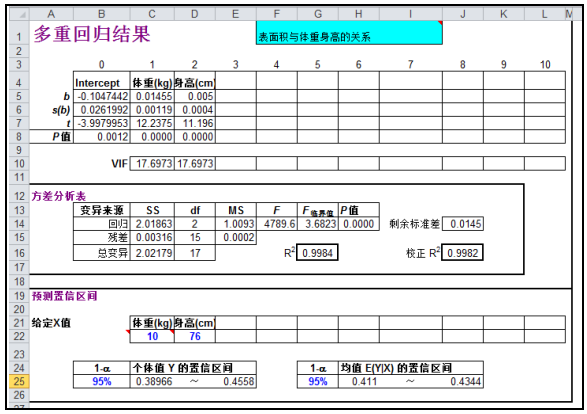


图 10-27 “回归”工作表

③在“残差”工作表（见图 10-28）中给出了残差图与正态概率图，用于判断资料是否满足多重回归的数据要求。残差图的纵轴为残差，横轴是源数据变量（包括自变量和应变量，可选），如果残差图散点均匀分布在 0 水平轴上下，表示数据满足要求，否则需采用数据变换等措施。正态概率图的横轴为残差，纵轴为残差对应的标准正态离差值 z，如果散点呈一条直线，则可认为 Y 变量近似服从正态分布。

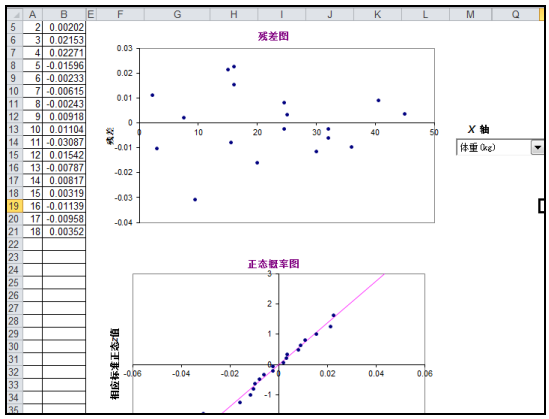


图 10-28 “残差”工作表

④“相关”工作表（见图 10-29）给出了两两变量之间的简单 Pearson 相关系数。这一结果与图 10-24 所示结果相同。



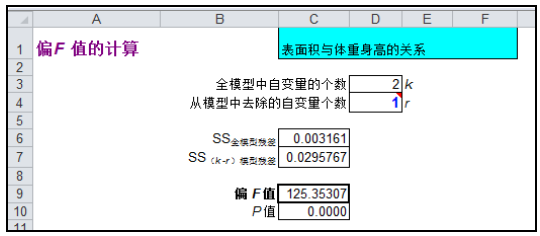
	A	B	C	D	E	F	G	H	I	J	K	L
1	相关系数矩阵				表面积与体重身高的关系							
2												
3			1	2	3	4	5	6	7	8	9	10
4			体重(kg)	身高(cm)								
5	1	体重(kg)	1.0000									
6	2	身高(cm)	0.9713	1.0000								
7	3											
8	4											
9	5											
10	6											
11	7											
12	8											
13	9											
14	10											
15												
16	Y 表面积(m2)		0.9927	0.9914								
17												

图 10-29 “相关”工作表

⑤“偏 F 值”工作表（见图 10-30）用于判断某变量或某些变量在回归方程中是否有作用，用于变量的筛选。如果有作用则应保留，无作用则应剔除。偏 F 值的计算公式为：

$$F_{偏}=\frac{\left(SS_{k\text{模型残差}}-SS_{(k\text{-}r)\text{模型残差}}\right)/r}{SS_{k\text{模型残差}}/(n-k-1)} \tag{10-18}$$

其中，k 模型是指所有变量均在模型中时的全模型，(k-r)模型是指剔除了“数据”工作表右侧 r 个自变量后，其余自变量与应变量之间的模型。如果偏 F 值足够大，即检验 P 值足够小，则 r 个自变量至少有一个自变量不可忽略；如果偏 F 值并非足够大，则这 r 个自变量便可从模型中剔除。本例去除变量“身高”后，得到 $F_{偏}=125.35$，P 值远小于 0.05，表示身高对体表面积的作用较大，从而身高对体表面积的作用不可忽视。



	A	B	C	D	E	F
1	偏F 值的计算		表面积与体重身高的关系			
2						
3		全模型中自变量的个数	2	k		
4		从模型中去除的自变量个数	1	r		
5						
6		$SS_{全模型残差}$	0.003161			
7		$SS_{(k-r)\ 模型残差}$	0.0295767			
8						
9		偏 F 值	125.35307			
10		P 值	0.0000			
11						

图 10-30 “偏 F 值”工作表

温馨提示：用户可将变量数据输入到“数据”工作表单元格区域 B5:B204，以及 D5:M204 范围内，改变“回归”工作表的“给定 X 值”及置信度，选取“残差”工作表的残差图“X 轴”，输入“偏 F 值”工作表的“从模型中去除变量个数”，便可获得所需结果。

10.5 加权 logistic 回归

二分类变量在生活中经常遇到，如考试及格与不及格、某种疾病在发病 1 年后死亡与

不死亡、做某一件事的结果为成功与失败。为了分析二分类变量与其他变量之间的关系，通常可采用 logistic 回归分析。二分类变量的结果可采用率 P 表示，如及格率、死亡率、成功率等，如果 n 个体中有 s 个体成功，则成功率 $P=s/n$。为了分析自变量 X（如年龄）是否对成功率 P 有影响，可拟合 logistic 回归方程，即

$$\text{logit}(\hat{p}_i)=\ln\frac{\hat{p}_i}{1-\hat{p}_i}=\hat{Y}=b_0+bX_i \tag{10-19}$$

其对应的总体方程可写为：

$$\text{logit}(\pi_i)=\ln\frac{\pi_i}{1-\pi_i}=\beta_0+\beta X_i+\varepsilon \tag{10-20}$$

这里 π_i 为样本率 p_i 的总体率，β_0、β 分别为样本回归系数 b_0、b 的总体回归系数，ε 为总体残差。

公式（10-19）也可写为：

$$\hat{p}_i=\frac{1}{1+\mathrm{e}^{-(b_0+bX_i)}} \tag{10-21}$$

如果给定自变量 X 值，可由公式（10-21）预测事件的发生率。

回归系数 b_0、b 的求解方法有多种，SAS、SPSS 等主流软件主要采用最大似然法进行估计，本章仅采用加权最小二乘法进行回归系数估计。计算公式为：

$$b=\frac{l_{wxy}}{l_{wxx}}=\frac{\left(\sum WXY-\dfrac{\sum WX\sum WY}{\sum W}\right)}{\left(\sum WX^2-\dfrac{\left(\sum WX\right)^2}{\sum W}\right)} \tag{10-22}$$

$$b_0=\frac{\sum WY-b\sum WX}{\sum W} \tag{10-23}$$

其中 $Y_i=\ln\dfrac{p_i}{1-p_i}$ 为率 p_i 的 logit 变换值，$W_i=n_ip_i\left(1-p_i\right)$ 为率 p_i 的方差，以 W_i 作为加权法的权重。由公式

$$OR\ =\ \exp(b) \tag{10-24}$$

计算变量 X 对率影响的优势比。

例 10-8：为了研究年龄是否与食管癌有关，下面对图 10-31 中数据作 logistic 回归分析。

解：按上述介绍方法，已编制好 logistic 回归求解程序（见光盘文件“10-16 加权 logistic

回归.xlsx”)。

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Logistic回归（加权最小二乘法）												
2													
3		X	n	s		p	Y	W	WX	WY	WXY	WX^2	WY^2
4	编号	年龄（岁）	研究个体数	病例数		s/n	$\ln(p/(1-p))$	$np(1-p)$					
5	1	30	10	1		0.1	-2.1972246	0.9	27	-1.9775	-59.3251	810	4.345016
6	2	30	106	0		1E-10	-23.025851	1.06E-08	3.18E-07	-2.4E-07	-7.3E-06	9.54E-06	5.62E-06
7	3	40	30	4		0.133333	-1.8718022	3.466667	138.6667	-6.48891	-259.557	5546.667	12.14596
8	4	40	169	5		0.029586	-3.4904285	4.852071	194.0828	-16.9358	-677.432	7763.314	59.11322
9	5	50	54	25		0.462963	-0.14842	13.42593	671.2963	-1.99268	-99.6338	33564.81	0.295753
10	6	50	159	21		0.132075	-1.8827312	18.22642	911.3208	-34.3154	-1715.77	45566.04	64.60675
11	7	60	69	42		0.608696	0.44183275	16.43478	986.087	7.261425	435.6855	59165.22	3.208335
12	8	60	173	34		0.196532	-1.4081134	27.31792	1639.075	-38.4667	-2308	98344.51	54.16552
13	9	70	37	19		0.513514	0.05406722	9.243243	647.027	0.499756	34.98295	45291.89	0.02702
14	10	70	124	36		0.290323	-0.8938179	25.54839	1788.387	-22.8356	-1598.49	125187.1	20.41087
15	11	80	5	5		1	23.0258508	5E-10	4E-08	1.15E-08	9.21E-07	3.2E-06	2.65E-07
16	12	80	39	8		0.205128	-1.3545457	6.358974	508.7179	-8.61352	-689.082	40697.44	11.66741

图 10-31　logistic 回归数据与部分中间计算结果

①单元格区域 B4:D4 供用户输入变量名，单元格区域 B5:D204 供用户输入变量名对应的数据。第 F 列计算了样本率 p_i（为了避免出现样本率为 0 或 1 的情况，当样本率=0 时则令等于10^{-10}，当样本率=1 时则令等于 $1-10^{-10}$），第 G、H 列分别计算了 logit 变换值 Y_i 与权重 W_i，第 I 至 M 列为计算公式（10-22）、公式（10-23）作准备。

②在图 10-33 中，单元格区域 P4:W4，是对图 10-31 中第 C 列以及第 G 至 M 列数据求和，目的是为计算公式（10-22）、公式（10-23）作准备。由这两个公式计算得到常数项 b_0=-3.0525、回归系数 b=0.0346。由此可由公式（10-21）得到：

$$\hat{p}_i = \frac{1}{1+\mathrm{e}^{3.0525-0.0346X}} \qquad (10\text{-}25)$$

第 O 列以及单元格 R12 就是根据此公式计算得到的。根据公式(10-24)获得 OR=1.0352，说明年龄每增加 1 岁，食管癌的优势增加 3.52%；年龄每增加 10 岁，食管癌的优势增加 41.37%（$OR=\exp(10\times b)=\exp(10\times 0.0346)=1.4137$）。

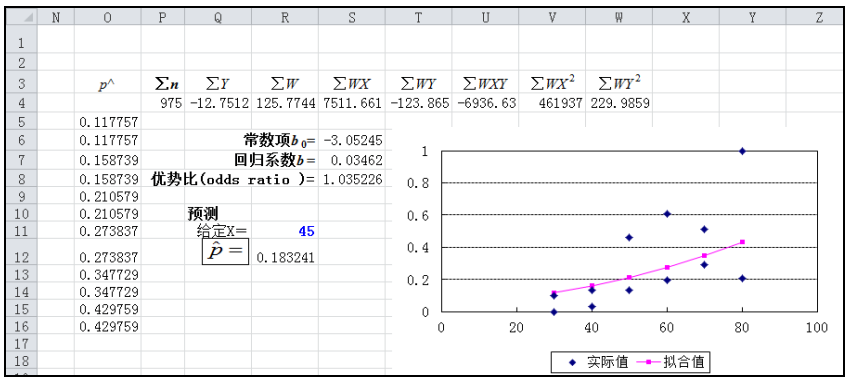


图 10-32　logistic 回归计算结果

温馨提示：该程序可处理 200 个观察个体，一个自变量。编写此程序的目的是简单介绍 logistic 回归和加权最小二乘法。用户只需在单元格区域 B5:D204 中输入相应数据，便可获得图 10-32 中的有关结果。为了给定某一自变量值对样本率进行预测，可在单元格 R11 中改变 X 值。

本章既给出了函数编写程序，也给出了 Excel 自带工具实现相关与回归。这两种方法各有其特点，读者既可以单独使用，也可以同时使用这两种方法。

高级应用篇

第 11 章　变量求解与方案优选

第 12 章　预测分析

第 13 章　寿命表

第 14 章　随机化分组与样本含量估计

第 15 章　高级科研统计图与 Meta 分析

第 16 章　诊断试验评价

第11章

变量求解与方案优选

Excel“数据→数据工具→模拟分析”命令之中含有“单变量求解”、“方案管理器”，在“数据→分析”之中含有“规划求解”（如果没有，则需单击“文件→选项→加载项”按钮进行安装）。首先介绍这几个数据分析工具，然后介绍用 Excel 实现几种方案优选的方法。此外，本章的最后一节也将介绍利用多个现有指标对评价对象进行综合评价的方法。

11.1　单变量求解

单变量求解相当于数学上的求解反函数，即对于 $y=f(x)$，给定 y 值，反过来求解 x。一般情况下，可按 y 与 x 之间的依赖关系，构造一个反函数 $x=\varphi(y)$，但当变量之间的依赖关系较为复杂（如非线性函数）时，构造反函数的工作往往相当困难，而利用功能区中的“数据→模拟分析→单变量求解”命令及 $y=f(x)$ 函数形式可方便地完成反函数的计算。

11.1.1　目标搜索

Excel 的单变量求解是一种目标搜索技术，如果将某变量存放在“可变单元格”中，含有该变量的公式存放在“目标单元格”中，Excel 不断改变指定“目标单元格”中的数值，使该公式得出给定的“目标值”，从而求解该变量值。

例 11-1：已知某医药公司年药品销售毛收入为 1385220 元，药品成本占毛收入的 35%，销售活动费用占毛收入的 5%，销售税金及其他开销占毛收入的 28%，根据公式：

利润=毛收入–药品成本–销售活动费–税金及其他

=毛收入–35%×毛收入–5%×毛收入–28%×毛收入

=（1–35%–5%–28%）×毛收入

=32%×毛收入

求得利润=32%×1385220=443270.40 元。问：如果要使年利润达到 450000 元，年销售毛收入应达到多少？

解：单变量求解可按如下步骤进行。

① 按图 11-1a 所示在单元格区域 A1:B7 中键入文字、数据和公式（输入后显示结果如图 11-1b 所示）（见光盘文件“11-1 变量求解.xlsx”）。单元格区域 B4:B7 中的 Excel 公式由所给已知条件确定。

	A	B
1	药品销售	
2		
3	毛收入	1385220
4	药品成本	=35%*B3
5	销售活动费用	=5%*B3
6	税金及其它	=28%*B3
7	利润	=B3-B4-B5-B6

图 11-1a　计算公式

	A	B
1	药品销售	
2		
3	毛收入	1385220.00
4	药品成本	484827.00
5	销售活动费用	69261.00
6	税金及其它	387861.60
7	利润	443270.40

图 11-1b　显示数据

② 单击“数据→数据工具→模拟分析→单变量求解”命令弹出“单变量求解”对话框，如图 11-2 所示。在“目标单元格”编辑框内选取单元格 B7（利润），该单元格是待求解公

式所在位置；在“目标值”编辑框中键入需要得到的目标值 450000；在“可变单元格”编辑框中选取单元格 B3，该单元格为待调整数值（毛收入）所在位置，此单元格必须直接或间接被“目标单元格”中的公式所引用。

③单击“确定”按钮，将弹出“单变量求解状态”对话框，显示已求得解答，且当前解与目标值相同（如图 11-3 所示）。



图 11-2 “单变量求解”对话框

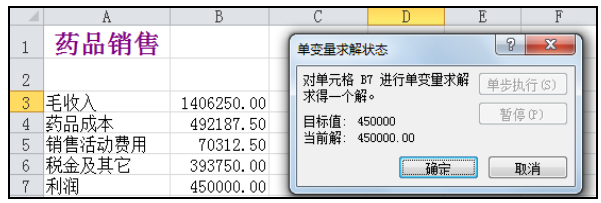


图 11-3 单变量求解状态及结果

④继续单击“确定”按钮，得到求解结果。由此可见该医药公司销售毛收入需要增加到 1406250 元，才能达到预定目标（利润达到 45 万元）。

11.1.2 求解非线性方程

数学上求解非线性方程时通常采用叠代法、割线法等，手工计算相当困难，而利用 Excel 的单变量求解命令求解方程则相当方便。

例 11-2：求解非线性方程 $4x^4+5x^3-7x^2+9x=100$。

解：单变量求解非线性方程可按如下步骤进行。

①选定单元格 B1，并在其中键入公式“=4*A1^4+5*A1^3-7*A1^2+9*A1”。令单元格 A1 存放 x 的解，因为单元格 A1 中没有输入任何值（是空的），所以在单元格 B1 中输入完公式后，公式将单元格 A1 的值默认为 0，由公式计算单元格 B1 获得的值也为 0。

②单击“数据→数据工具→模拟分析→单变量求解”命令，弹出“单变量求解”对话框，在“目标单元格”编辑框内选取单元格 B1，在“目标值”编辑框中键入目标值 100，指定“可变单元格”为 A1，单击“确定”按钮，如图 11-4 所示。由此可见，获得变量 x 的解为 2.0346，此时单元格 B1 并非精确等于 100，而是获得了一个接近 100 的数，即 100.0000672。

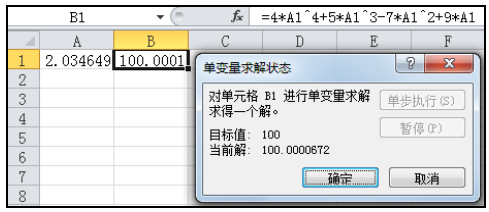


图 11-4 单变量求解状态及结果

11.2　规划求解

Excel 中的“规划求解”不仅具有上一节介绍的“单变量求解”功能，而且还可以在给定或不给定约束条件下求极值（即最大值或最小值）。在第 10 章的 10.3 节已利用“规划求解”求解非线性回归系数，这里将再列举两个例子，详细说明这一方法。

例 11-3：已知非线性方程 $y=4x^4+5x^3-7x^2+9x-100$。①求 $y=0$ 时 x 的解；②求 y 的极值及相应的 x 值。

解：①求 $y=0$ 时的 x 值。

对于这个问题，实际上已在例 11-2 中利用“单变量求解”求得，这里再利用“规划求解”来求 x 值，以便使读者初步了解“规划求解”的内容。

① 在单元格 B1 中键入公式“=4*A1^4+5*A1^3–7*A1^2+9*A1–100”，令单元格 A1 中存放 x 的解。

② 单击功能区“数据→分析→规划求解”命令，在“规划求解参数”对话框中，设置“设置目标”为“$B$1”（用鼠标选择）；选择“目标值”，并在空白框中填入数字“0”；用鼠标选择“通过更改可变单元格”为“$A$1”（如图 11-5 所示）。单击“求解”按钮，将弹出“规划求解结果”对话框（如图 11-6 所示），再单击“确定”按钮，可获得 $x=2.034649$，这一结果与“单变量求解”获得的结果相同（如图 11-4 所示）。

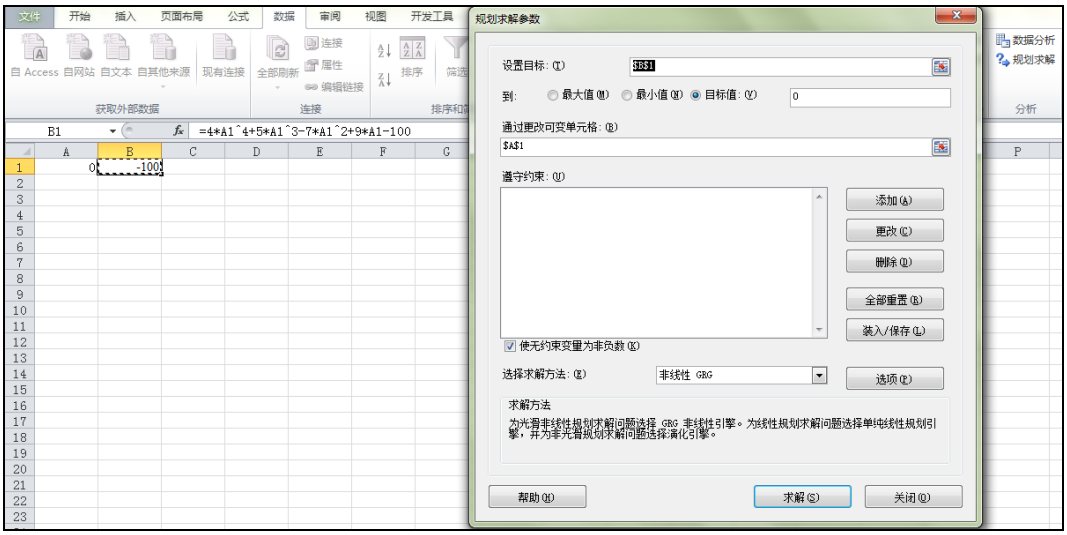


图 11-5　“规划求解参数”对话框

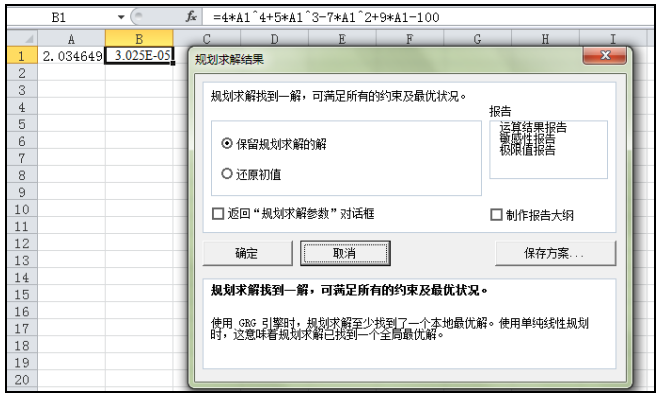


图 11-6 “规划求解结果”对话框

②求 y 的极值及相应的 x 值。

为了获得极值及其对应的 x 值，只需在图 11-5 中选择最大值或最小值即可。对于例 11-3，在其他条件不变的情况下（单元格 A1 的值为“0”），“到：”后选择“最大值”，单击“求解”按钮，获得最大值为 3.323E+31，相应 x 为 53687091，并在“规划求解结果”对话框下方提示“目标单元格的值未收敛”；“到：”后选择“最小值”，单击“求解”按钮，获得最小值为 –100，相应 x 为 0，并在下方提示“规划求解找到一解，可满足所有的约束及最优状况”。如果去掉图 11-5 中的“使无约束变量为非负数”，则得到最小值为–126.7284，相应 x 为–1.6656，并在下方提示“规划求解收敛于当前的解，可满足所有的约束及最优状况”。

注意：单元格 A1 中不同的初始值有可能会得到不同的最大值或最小值，如例 11-3 的最大值求解。

由以上例子可见，“规划求解”是对直接或间接与目标单元格中公式相联系的可变单元格中的数值进行调整，最终利用目标单元格公式求得期望的结果。

在图 11-5 的“通过更改可变单元格”编辑框中，键入每个“可变单元格”的名称或引用，用英文逗号分隔不相邻的引用，“可变单元格”必须直接或间接与“设置目标”的单元格相联系，最多可以指定 200 个“可变单元格”。

在图 11-5 的“遵守约束”列表框右侧，单击“添加”按钮，弹出如图 11-7 所示的“添加约束”对话框，由此可设定规划求解的约束条件，“单元格引用”与“约束”之间可采用 <=、=、>=、int 或 bin 等关系符号，这些关系符号分别表示小于等于、等于、大于等于、整数、二进制。

单击图 11-5 中的“选项”按钮，弹出“选项”对话框（如图 11-8 所示）。在此对话框中，可以设置规划求解过程的一些高级属性，如精确度、求极限值、演化等，每一选项都有默认设置，可以满足大多数情况下的要求。要深入了解这些选项，可单击“规划求解”

对话框中的“帮助”按钮。

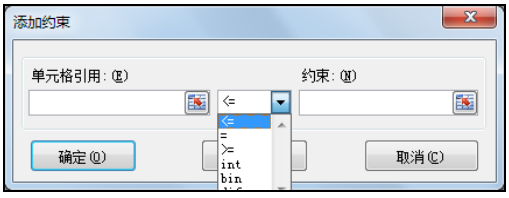


图 11-7　“添加约束”对话框

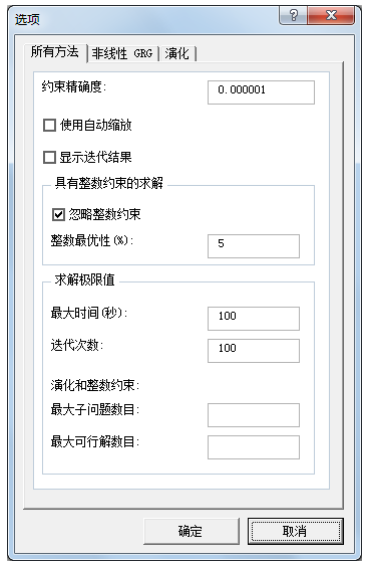


图 11-8　规划求解“选项”对话框

例 11-4：某药厂专门生产药品原材料，计划年产两种药品原材料分别为 x_1、x_2 吨，此问题的数学模型为：$y=7x_1+11x_2-x_1^2-2x_2^2$；约束条件为：$5x_1+2x_2\leqslant 8$，$x_1\geqslant 0$，$x_2\geqslant 0$。

问：为了获得最大利润，生产这两种药的原材料各需多少吨？

解：这是一个带约束条件的求 y 极大值的问题。具体求解步骤如下：

① 令单元格 D2、E2 分别为需要求解的两种药的原材料吨数 x_1、x_2，在单元格 A2 中键入公式“=7*D2+11*E2+(–1)*D2*D2+(–2)*E2*E2”，以便建立数学模型。

② 在单元格 B2 中键入公式“=5*D2+2*E2”，以便建立约束条件。

③ 单击“数据→分析→规划求解”命令，在“规划求解”对话框中，设置“设置目标”为$A$2（用鼠标选择）；“到：”右边选择“最大值”，设置“通过更改可变单元格”为$D$2:$E$2。在“遵守约束”下方根据约束条件添加“约束”。单击“添加”按钮，在“添加约束”对话框左侧的“单元格引用位置”框内引用单元格$B$2，中间选择“<=”，在右侧“约束值“框内填写 8，由此建立约束条件 1 为“$B$2<=8”。再单击“添加”按钮，建立约束条件 2 为“$D$2>=0”；继续再单击“添加”按钮，建立约束条件 3 为“$E$2>=0”，单击“确定”按钮，此时的“规划求解参数”对话框如图 11-9 所示（参见光盘文件“11-2 约束极值规划求解.xlsx”）。

④ 单击“求解”按钮，屏幕上弹出“规划求解结果”对话框，选中“保留规划求解的

解”（默认）项，单击“确定”按钮，得到所需结果。即该药厂可年生产第一种药品原材料 0.7 吨、第二种药品原材料 2.2 吨，并获得最大利润 19.0 万元。

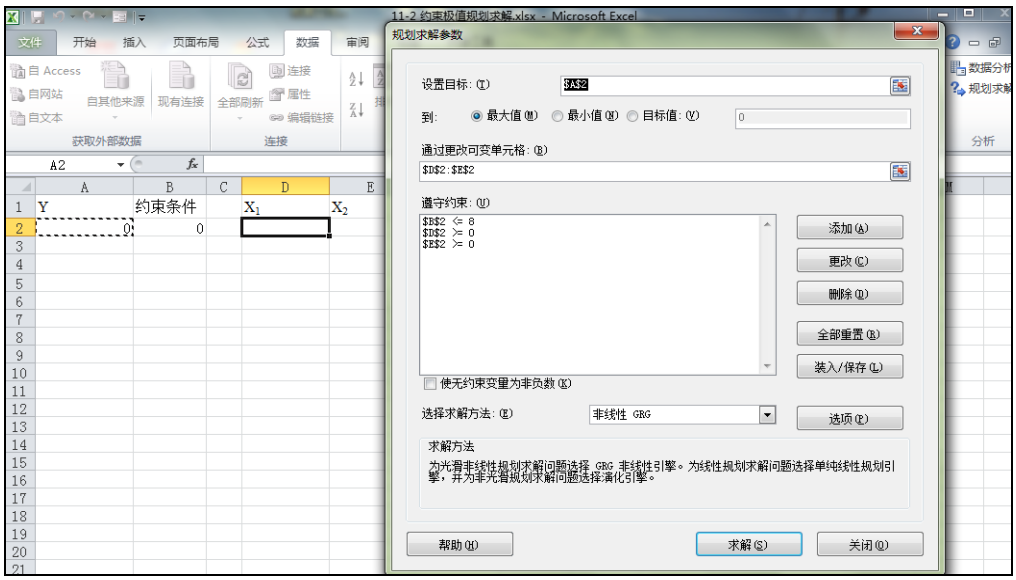


图 11-9　建立约束条件后的“规划求解参数”对话框

11.3　方案管理器

在 11.1 节中的“单变量求解”只考虑了一个可变单元格，对于一些复杂的问题，常常需要考查更多的情况，如医药公司为了获得更高利润，既可以通过增加销售收入的方案获得高利润，也可以在销售收入不变的情况下，通过降低药品成本及销售活动费的方案获得高利润。Excel 提供的“方案管理器”可模拟出各种方案的结果，通过对比分析不同方案的优劣，可从中选择最佳方案。

例 11-5：对于例 11-1 中的数据设定两种方案。方案 1 是增加毛收入至 1400000 元；方案 2 是在销售毛收入不变的情况下，降低药品成本至 450000 元，降低销售活动费用至 65000 元（利润=毛收入–药品成本–销售活动费–税金及其他）。请比较哪种方案更可行？

解：具体可按下列步骤进行。

1. 创建方案

创建方案是方案分析的关键，应根据实际问题的需要和可行性创建一组方案。在创建方案之前，为了使创建的方案能够明确地显示有关变量，以及为了将来进行方案总结时便

于阅读，需要先给有关变量所在的单元格命名。其具体操作步骤如下：

①打开光盘文件“11-3 方案.xlsx”，单击“公式→定义的名称→名称管理器”命令，在弹出的“名称管理器”对话框中单击“新建”按钮，再在弹出的“新建名称”对话框中，依次键入名称（如将毛收入的数据(B3)命名为“销售收入”）和引用位置（如图 11-10a 所示）。

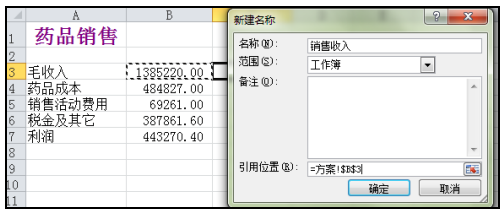


图 11-10a　新建名称

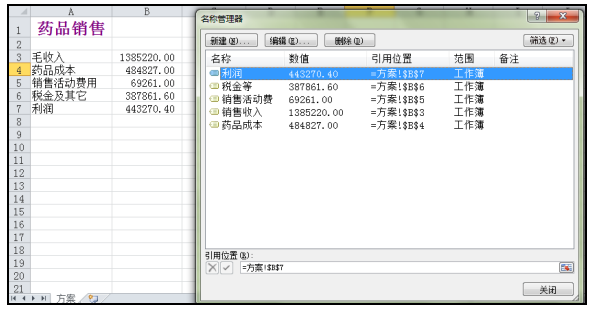


图 11-10b　名称管理器

②将 B3～B7 的 5 个单元格数据全命名完后，再单击“名称管理器”命令，会有如图 11-10b 所示的界面。此时可以对这 5 个命名进行编辑或删除处理（见光盘文件“11-3 方案.xlsx”）。

这时可按下述步骤逐个创建所需的方案。

①单击功能区“数据→数据工具→模拟分析→方案管理器”按钮，弹出“方案管理器”对话框，由于现在还没有任何方案，所以“方案管理器”对话框中间显示“未定义方案”的信息（如图 11-11 所示）。根据提示，单击“添加”按钮，出现“编辑方案”对话框。在“方案名”框中键入方案的名称，这里键入“增加收入”。然后指定销售收入所在的单元格 B3 为“可变单元格”（如图 11-12 所示），单击“确定”按钮。

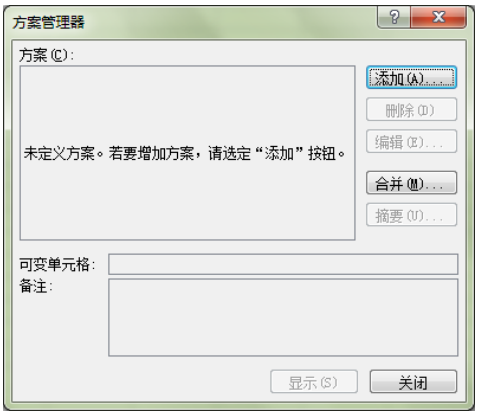


图 11-11　“方案管理器”对话框

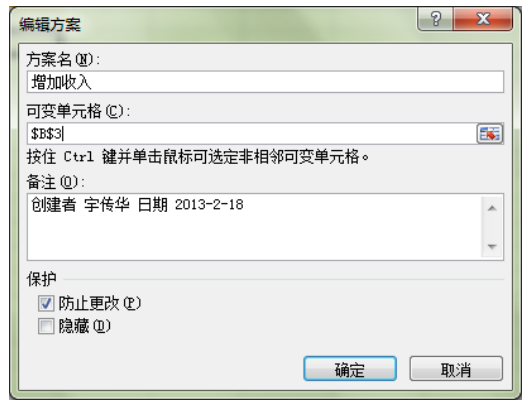


图 11-12　增加收入“编辑方案”对话框

②在弹出的“方案变量值”对话框中键入 1400000（如图 11-13 所示），单击“确定”按钮。“增加收入”方案创建完毕，相应的方案自动添加到方案管理器的方案列表中。

按照上述步骤再建立“降低成本”方案（注意此方案的可变单元格为 B4:B5 两个，按“确定”后，分别填充值 450000、65000）。这时的“方案管理器”对话框如图 11-14 所示。



图 11-13　键入方案 1 的变量值

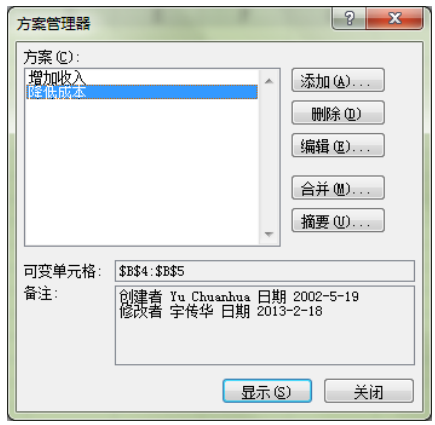


图 11-14　方案建立后的方案管理器

2. 方案摘要

单击图 11-14 中的“摘要”按钮，将弹出如图 11-15 所示的“方案摘要”对话框。在此对话框的“结果单元格”框中选取单元格 B7，“报表类型”选择“方案摘要”（默认），单击“确定”按钮，获得的方案总结结果如图 11-16 所示。

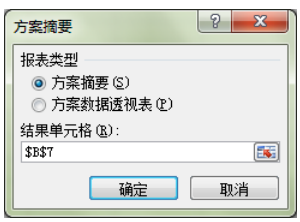


图 11-15　“方案摘要”对话框

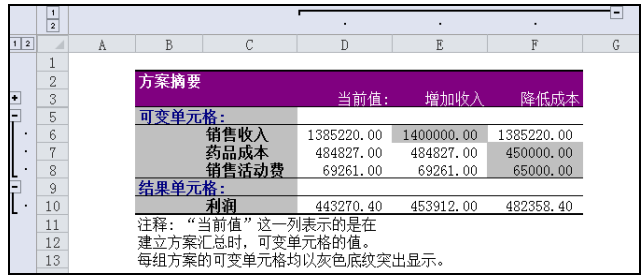


图 11-16　方案总结结果

在方案总结结果的方案摘要中，“当前值”一列显示的是建立方案汇总前的可变单元格原始数值，每组方案的可变单元格均以灰色底纹突出显示，根据各个方案模拟数据计算出的目标值也同时显示在总结中，见单元格区域 D10:F10。比较两个方案的结果单元格“利润”值可见：“降低成本”将获得利润为 482358.40 元，“增加收入”将获得利润为 453912.00 元。因此，方案 2 即“降低成本”，对目标值的影响相对较大，说明方案 2 可行性更好。

如果选择图 11-15 中的“方案数据透视表”，则可按第 5 章的数据透视表形式显示方案的结果。

11.4　常见的方案优选方法

常见的方案优选方法有：确定型决策、非确定型决策和风险型决策三类。

11.4.1　确定型决策

所谓确定型决策是指决策的问题只存在一种自然状态，即未来的事件以及与事件有关的各种条件都是确定的，这时的决策比较简单，只需计算出各种条件下的成本、收益等指标，按照特定的目标从中选择最佳方案即可。

例 11-6：某材料可分别从甲、乙、丙三地采购，材料质量相同，材料费和运费见表 11-1。

表 11-1　向三地采购材料的价格表（元/吨）

	材 料 费	运　　费	合　　计
甲地	200	6.1	206.1
乙地	198	6.2	204.2
丙地	203	6.0	209.0

问：从何地采购最省钱？

解：通过计算合计，可简单挑选合计费用最少的乙地采购该材料。

为了能够处理更多方案的情况，以及便于当方案修改时能快捷地找到最佳方案，可以利用 Excel 的查找函数 LOOKUP 自动选择最佳方案。其具体操作步骤如下：

① 按图 11-17 所示将表 11-1 中的数据输入，并建立合计列公式求合计。

	A	B	C	D	E	F	G
1							
2	表　向三地采购材料地价格表（元/吨）					最佳方案	=LOOKUP(MIN(D4:D6)
3		材料费	运费	合计			
4	乙地	198	6.2	=B4+C4			
5	甲地	200	6.1	=B5+C5			
6	丙地	203	6	=B6+C6			

图 11-17　合计列公式

② 选定要显示最佳方案名称的单元格 G2，选择“公式→函数库→插入函数”命令，在函数分类列表框中选择“查找与引用”，在函数名列表框中选择“LOOKUP 函数”。在弹出的 LOOKUP 函数的“选定参数”对话框中选择“lookup_value, lookup_vector, result_vector”组

合方式（如图 11-18 所示）。

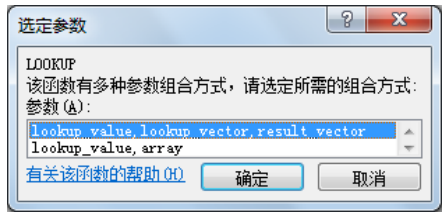


图 11-18　LOOKUP 函数的“选定参数”对话框

③在 LOOKUP 函数参数对话框的 Lookup_value 框内输入函数“MIN(D4:D6)”（如果取最大值则应选择 MAX 函数），该参数为要查找的数值；在 Lookup_vector 框内输入“D4:D6”，该参数为要查找的单元格区域；在 Result_vector 框内输入“A4:A6”，该参数为返回值所对应的单元格区域，如图 11-19 所示（见光盘文件“11-4 确定型决策.xlsx”）。单击“确定”按钮。

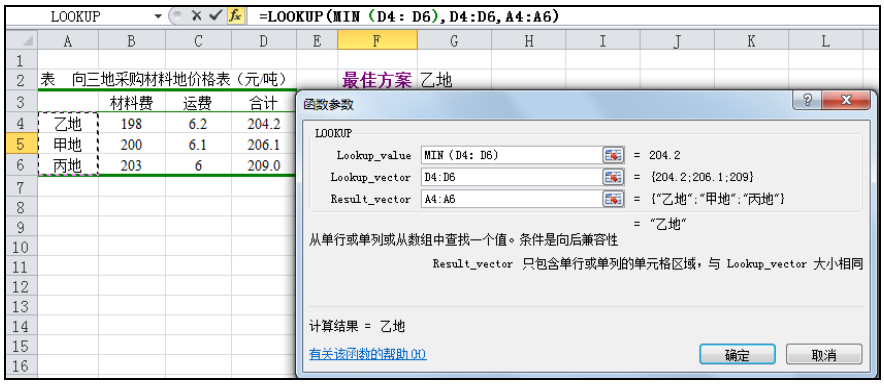


图 11-19　LOOKUP 函数参数对话框的参数填入

④选取单元格区域 A4:D6，按合计列从小到大的顺序排序，以免 LOOKUP 函数查找不到正确结果。

本例最佳方案为“乙地”，即从乙地采购材料最省钱。

11.4.2　非确定型决策

在实际工作中，通常遇到两种以上的自然状态，其出现的概率也无法预测，这种情况下作出的决策为非确定型决策，通常包括乐观、悲观、折中、遗憾值等原则决策。

例 11-7：某药厂生产药物有三个行动方案，分别为大量生产、中量生产和小量生产（分别记为 A1、A2、A3），出现的自然状态为疾病大流行、局部流行、散在发生（分别记

为 D1、D2、D3），它们出现的概率未知，其相应的收益值如表 11-2 所示，试据此选择行动方案。

表 11-2　某药厂生产药物三种自然状态收益值（万元）

方　案	自然状态		
	D1	D2	D3
A1	60	40	–20
A2	40	25	–10
A3	10	15	5

（1）乐观原则

所谓乐观原则是看好未来市场需求，在决策时总是选择收益最大或损失最小的方案，其基本思路是先由各方案计算出收益最大值，然后在每个方案的最大值中找出最大值，所以也称“大中取大法”。

具体计算方法如下：

将表 11-2 中的数据输入到工作表中，如图 11-20 所示（见光盘文件“11-5 乐观决策.xlsx”）。在第 E 列利用 MAX 函数计算出各方案在不同状态下的最大值，如在单元格 E3 中输入公式“=MAX(B3:D3)”，将此公式拖放至单元格 E5。对单元格区域 A3:E5 按第 E 列（最大值）排升序。在单元格 H2 中利用自动查找函数式“=LOOKUP(MAX(E3:E5),E3:E5,A3:A5)”，查找最大值所在列中最大者所对应的方案名（如图 11-20 所示）。按“乐观原则”本例选择“A1”方案，即选择大量生产该药的方案。

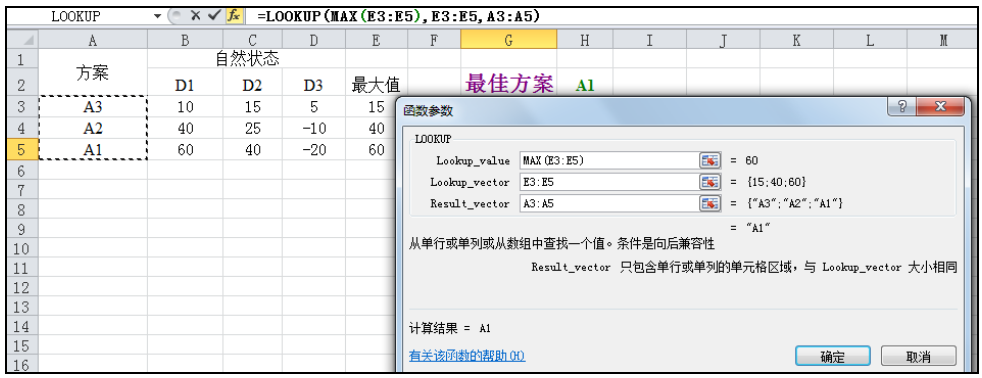


图 11-20　乐观原则决策

（2）悲观原则

所谓悲观原则，是决策时从最坏情况作打算，尽量减少风险。其基本思路是先由各方案计算出收益最小值，然后在每个方案的最小值中找出最大值，所以也称“小中取大法”。

具体计算方法如下：

在第 E 列利用 MIN 函数，计算出各方案在不同状态下的最小值，即在单元格 E3 中输入公式“=MIN(B3:D3)”，然后拖放此公式至 E5。在单元格 H2 中利用自动查找函数式“=LOOKUP(MAX(E3:E5),E3:E5,A3:A5)”，查找最小值所在列中最大者所对应的方案名（注意：对单元格区域 A3:E5 按第 E 列排升序），如图 11-21 所示（见光盘文件“11-6 悲观决策.xlsx”）。按“悲观原则”本例选择方案为 A3。

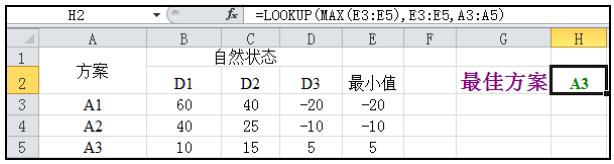


图 11-21　悲观原则决策

（3）折中原则

更多的情况下，决策时既不简单地根据乐观原则，也不完全按照悲观原则，而是采用介于二者之间的折中原则，这样既不过于冒险，也不过于保守。其基本思路是先由决策者主观地选取一个介于 0～1 之间的乐观系数α，由

折中收益值=α（最大收益值）+（1−α）（最小收益值）

获得每个方案的折中值，然后在每个方案的折中值中找出最大值。

具体计算方法如下：

在第 E 列利用 MAX 函数计算出各方案在不同状态下的最大值，在第 F 列利用 MIN 函数计算出各方案在不同状态下的最小值。设乐观系数α为 0.7，在第 G 列根据上述公式计算出各方案的折中值。在单元格 J2 中利用自动查找函数式“=LOOKUP(MAX(G3:G5), G3:G5, A3:A5)”，查找折中值所在列中最大者所对应的方案名（注意：对单元格区域 A3:G5 按折中值排升序），计算公式如图 11-22 所示（见光盘文件“11-7 折中决策.xlsx”）。本例在乐观系数α为 0.7 时选择方案为 A1。

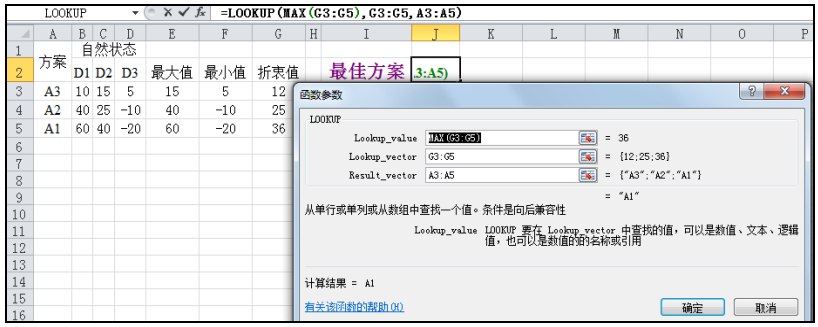


图 11-22　折中原则决策的公式

（4）遗憾值原则

所谓遗憾值原则，是决策时将每种状态的收益最高值作为该状态的理想目标，并将相应状态下其他值与理想值的差称作未达到理想值的遗憾值。其基本思路是先由各方案计算出理想值，再根据理想值计算出每个方案不同状态下的遗憾值矩阵，然后根据遗憾值矩阵计算出每个方案的最大遗憾值，最后在各最大遗憾值中选取最小者作为最佳方案，即根据遗憾值矩阵采用“大中取小法”选出最佳方案。具体计算步骤如下：

①利用 MAX 函数计算出每种自然状态的理想值。在单元格 F3 中输入公式“=MAX(B3:B5)”，然后将此公式向右拖放填充至 H3（如图 11-23 所示）。

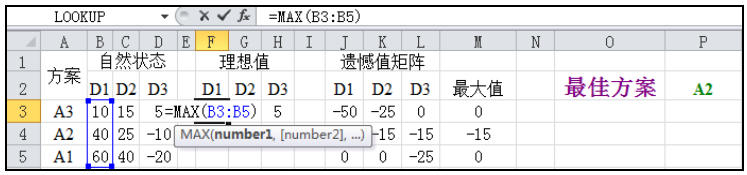


图 11-23　计算每种状态的理想值

②在工作表的另外一个区域 J3:L5 中建立遗憾值矩阵。在单元格 J3 中键入公式“=B3–F$3”，将此公式向右拖放填充至单元格 L3，然后选中单元格区域 J3:L3，再向下拖放填充至 L5。

③根据上述产生的遗憾值矩阵，按照大中取小原则选择方案。先在第 M 列计算出最大遗憾值，即在单元格 M3 中键入公式“=MAX(J3:L3)”，并将此公式拖放至 M5，然后再利用 LOOKUP 函数自动查找遗憾值一列中最小者所对应的方案名称，公式为“=LOOKUP(MIN(M3:M5),M3:M5,A3:A5)”。按遗憾值原则，本例选择方案为 A2，如图 11-24 所示（见光盘文件“11-8 遗憾决策.xlsx”）。

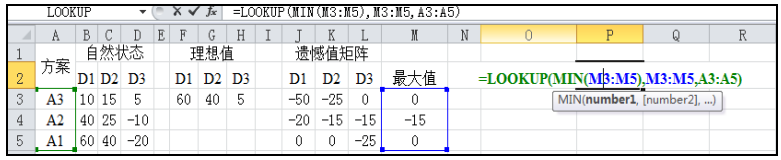


图 11-24　从遗憾值矩阵中按大中取小原则选择方案

从以上几个原则求解结果可见，非确定型决策对于同一问题，由于采用不同的决策原则，所得结果有所不同。究竟采用何种原则，取决于决策者对于各种自然状态的看法，因此，非确定型决策往往带有很大的主观随意性。

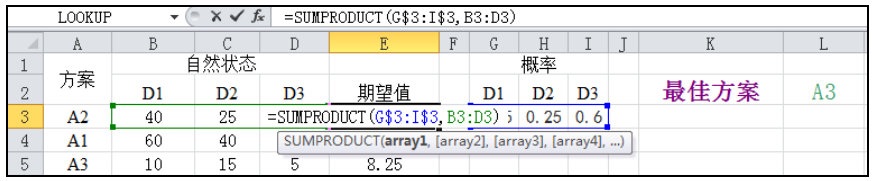
11.4.3　风险型决策

风险型决策与确定型决策和非确定型决策的不同之处在于，决策者虽然对于未来出现

哪种自然状态不能做出确定的判断，但能根据有关资料估计或计算出各种状态出现的概率。就例 11-7 来说，虽然该例不能确定某疾病在未来若干年是流行、局部流行还是散在发生，但如果根据有关资料或经验，可以知道这三种状态出现的概率分别为 0.15、0.25 和 0.60。此时的决策就不再是非确定型决策。这时无论做出哪种决定，都有一定的风险，所以这类决策问题称作风险型决策。

一般可按下列操作步骤作出决策。

① 计算收益值的期望值。对于某一方案，采用每种状态的概率乘以相应状态的收益值，然后相加获得。如图 11-25 所示（见光盘文件“11-9 风险决策.xlsx”），方案 A2 的期望值为：$0.15\times40+0.25\times25+0.6\times(-10)=6.25$。如果每种状态的概率存放在单元格区域 G3:I3，相应状态的收益值存放在单元格区域 B3:D5，单击单元格 E3，可在单元格 E3 中建立 Excel 公式“=SUMPRODUCT(G$3:I$3,B3:D3)”，求得方案 A2 的期望值。将 E3 中公式向下拖放至单元格 E5，可继续求得方案 A1、A3 的期望值。



	A	B	C	D	E	F	G	H	I	J	K	L
1	方案	自然状态					概率					
2		D1	D2	D3	期望值		D1	D2	D3		最佳方案	A3
3	A2	40	25	=SUMPRODUCT(G$3:I$3,B3:D3)			0.25	0.6				
4	A1	60	40									
5	A3	10	15	5	8.25							

图 11-25　风险型决策的期望值

② 从计算出的各方案的期望值中选取最大者。在单元格 L2 中键入公式“=LOOKUP(MAX(E3:E5),E3:E5,A3:A5)”，并对单元格区域 A3:E5 按第 E 列排序，采用风险型决策，获得例 11-7 的 D1、D2、D3 的状态概率分别为 0.15、0.25 和 0.60 时的最佳方案为 A3。

11.5　综合评价方法

为了总结以往工作，利用多个指标值得到评价对象的大小顺序的方法，称为综合评价（synthetic evaluation）法。综合评价方法很多，下面只介绍 TOPSIS 法和层次分析法。

11.5.1　TOPSIS 法

TOPSIS 法即逼近理想解排序法（Technique for Order Preference by Similarity to Ideal Solution，TOPSIS）。计算步骤如下：

① 同趋势化，即均变成越大越优的、同趋势的 y_{ij}，如果为低优指标或中性指标则需要取其转换值。

$$y_{ij}=\begin{cases}x_{ij} & \text{（高优指标）}\\ 1/x_{ij} & \text{（低优绝对数指标）}\\ 1-x_{ij} & \text{（低优相对数指标）}\\ a_j-\left|x_{ij}-a_j\right| & \text{（中性指标，}a_j\text{为最优值）}\end{cases} \tag{11-1}$$

其中 x_{ij} 为第 i 评价对象，第 j 指标的个体值，y_{ij} 为转换值，a_j 为第 j 指标的最优值。相对数是指率、构成比与相对比，其他为绝对数。

② 指标归一化。令 a_{ij}（与前面的中性指标最优值 a_j 含有不同）归一化值，采用公式：

$$a_{ij}=\frac{y_{ij}}{\sqrt{\sum_{i=1}^{n}y_{ij}^2}},\quad i=1,2,\cdots,n,\quad j=1,2,\cdots,m \tag{11-2}$$

对每一个体值进行变换。

③ 获得现有评价对象的第 j 指标的 a_{ij} 最大值 $a_{j\max}$ 与最小值 $a_{j\min}$。

④ 分别计算各评价对象的最优方案欧氏距离 D_i^+ 与最劣方案欧氏距离 D_i^-，即

$$D_i^+=\sqrt{\sum_{j=1}^{m}\left[w_j\left(a_{ij}-a_{j\max}\right)\right]^2},\quad D_i^-=\sqrt{\sum_{j=1}^{m}\left[w_j\left(a_{ij}-a_{j\min}\right)\right]^2} \tag{11-3}$$

其中 w_j 为每一个指标所占权重，$\sum w_j$=1。

⑤ 计算各评价对象与最优方案的相对接近程度：

$$C_i=\frac{D_i^-}{D_i^++D_i^-} \tag{11-4}$$

C_i 在 0 与 1 之间取值，越接近 1，表示该评价对象越接近最优水平；反之，越接近 0，表示该评价对象越接近最劣水平。

⑥ 按 C_i 大小将各评价对象排序，C_i 值越大，表示综合效益越好。

例 11-8：广西某医院收集了 2001～2005 年医院诊疗工作的 8 个指标数（如表 11-3 所示，参见胡灼君，同济医学院硕士论文，2007），试根据这些指标对该医院 5 年的诊疗情况作出评价。

解：“院内感染率”、“住院天数”值越低越好，是低优指标，其他 6 个指标为高优指标。“院内感染率”是相对数，采用“100%-感染率”；住院天数是绝对数，对齐取倒数（见图 11-26 的第 J 列和第 M 列），其他 6 个指标值保持不变，将其输入到光盘文件“11-10 TOPSIS.xlsx”的单元格区域 H4:AK103 中（如图 11-26 所示），程序会自动获得所需结果（见图 11-26 的第 A 至 E 列）。

结果分析：依照 C_i 对各评价对象进行排序，得到 2005 年诊疗情况最优，其次是 2001

年，排在第 3、4、5 的分别是 2004、2002、2003 年。

表 11-3　广西某医院 2001～2005 年的医院诊疗情况

年份	治愈率%	抢救	院内	诊断	病床	住院天数	职工平均负荷	
		成功率%	感染率%	符合率%	使用率%		门诊人数	住院人数
2001	85.76	72.1	3.55	87.16	81.76	12.86	575.17	151.22
2002	85.22	75.6	3.60	80.72	79.66	11.92	553.94	149.30
2003	85.74	76.4	3.37	85.91	70.78	11.92	584.79	136.37
2004	86.48	78.2	2.90	87.83	72.07	11.53	550.17	146.92
2005	86.96	79.4	3.18	92.60	74.81	11.69	758.28	154.93

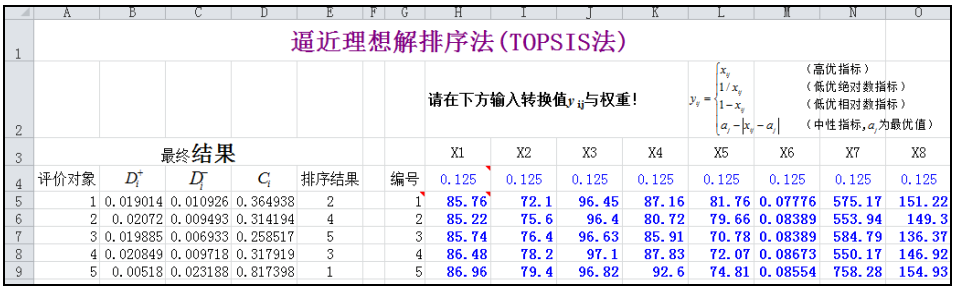


图 11-26　TOPSIS 法界面

温馨提示：该程序设置最大可适应于 100 个评价对象，30 个高优指标。用户只需修改图 11-26 中单元格区域 H5:AK105 中的数据，便可获得评价对象的排序结果。

如果可以获得每一评价指标的权重，则可实现加权 TOPSIS 法。例如，例 11-8 的 8 个指标的权重确定为（0.14、0.14、0.18、0.13、0.12、0.12、0.08、0.09），将这些权重在单元格区域 H4:AK4 中的相应位置键入，获得结果如图 11-27 所示（见光盘文件“11-10 TOPSIS1.xlsx”）。其排列顺序与相同权重相同，但接近度值相差更悬殊。

另一实例见光盘文件“11-10 TOPSISjll.xlsx”。

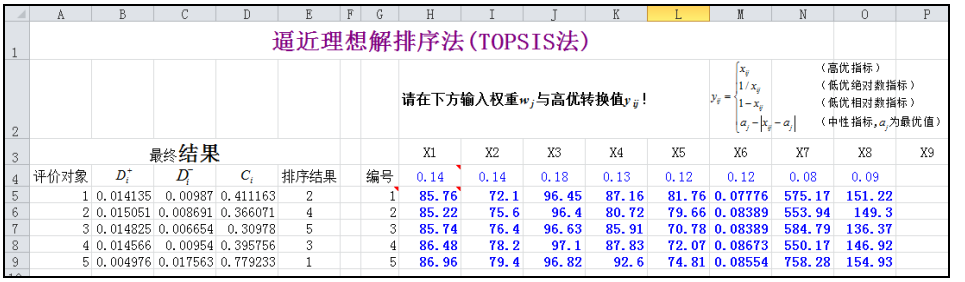


图 11-27　加权 TOPSIS 法界面

11.5.2　层次分析法

TOPSIS 法一般无权重，即使有权重也只是在同一层次进行处理（见公式 11-3）；而层次分析则是将指标分成若干层，再进行权重分配的方法。层次分析法（Analytic Hierarchy Process，AHP）由 Saaty TL 于 20 世纪 70 年代提出，是对总评价指标（树根）进行连续性分解，得到各级（各层）评价指标，并以最下层指标（树叶）作为衡量目标达到程度的评价指标，然后依据这些指标计算出综合评分，依综合评分大小来确定评价对象的优劣等级。

例 11-9：广西某医院收集了 2001～2005 年 5 类指标（诊疗工作、经营开发、科研能力、满意度、收费情况），并分别赋予权重为 0.40、0.32、0.1、0.1、0.08（相加为 1）。诊疗工作指标 8 个、经营开发指标 7 个、科研能力指标 6 个、收费情况指标 3 个的权重（相加均为 1），如图 11-28 所示（参见胡灼君，同济医学院硕士论文，2007），试对该医院 5 年的医院工作绩效作出评价。

解：计算步骤如下。

① 根据专家打分等方法获得各层次的权重。

② 乘积法求各层次评价指标的组合权重，如图 11-28 所示的最后 1 列，注意这些终点（树叶）组合指标权重之和一定为 1。

③ 将低优指标取倒数转化为高优指标。

④ 将每一高优指标除以该指标的最大值，获得无量纲指标 a_{ij}，为了避免 0 为分母的情况，应将最大值加上一个十分小的数（如 10E-10）。

⑤ 根据公式：

$$GI_i = \sum_{j=1}^{m} w_j \cdot a_{ij} \qquad (11\text{-}4)$$

求得每个评价对象的综合评分。

⑥ 根据综合评分大小进行优劣等级排序。

前三步采用手工完成，后三步可借助于光盘文件“11-11 AHP.xlsx”完成。

打开光盘文件“11-11 AHP.xlsx”，将所有高优指标（对低优指标取倒数转化为高优指标）的权重依次输入到单元格区域 F4:BC4，其相应指标值输入到单元格区域 F5:BC104 范围内（见光盘文件“11-11AHP.xlsx”），X1～X25 分别依次对应图 11-28 中的 25 个指标。程序便自动得到如图 11-29 所示的评价结果。

结果分析：依据综合评分对各评价对象进行排序，得到 2005 年医院工作绩效最优，2001、2002、2003、2004 年分别排第 2、3、4、5。

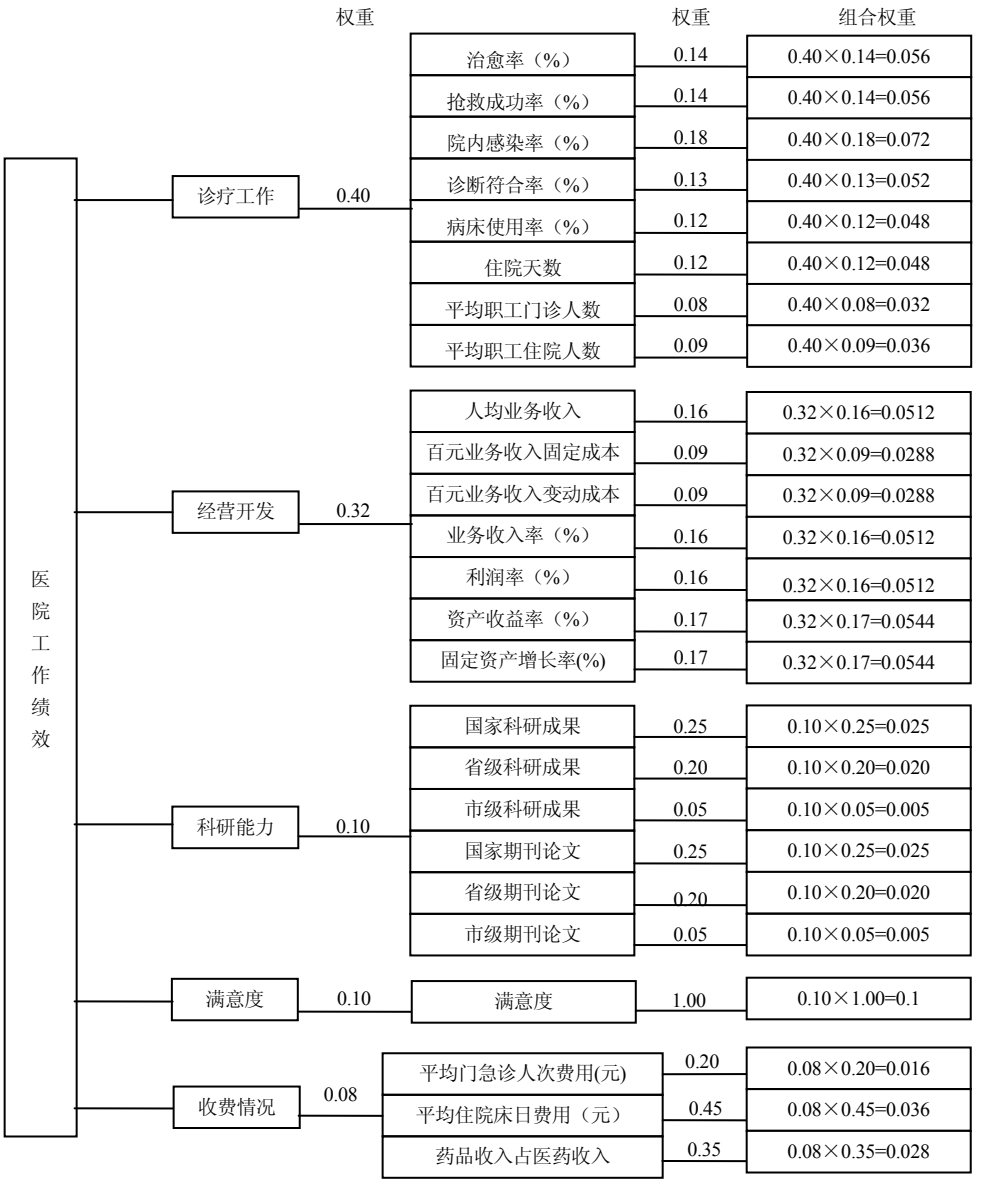


图 11-28　乘积法求各层次评价指标的组合权重

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	层次分析法（AHP， Analytic Hierarchy Process）													
2					1	请在下方输入高优指标值与权重！（如果非高优，则取其倒数使之转换为高优指标）								
3	最终结果					X1	X2	X3	X4	X5	X6	X7	X8	X9
4	评价对象	综合评分	排序结果		权重	0.056	0.056	0.072	0.052	0.048	0.048	0.032	0.036	0.0512
5	1	0.861593	2		1	85.76	72.1	0.28169	87.16	81.76	0.07776	575.17	151.22	53.41
6	2	0.80574	3		2	85.22	75.6	0.27778	80.72	79.66	0.08389	553.94	149.3	54.69
7	3	0.791444	4		3	85.74	76.4	0.29674	85.91	70.78	0.08389	584.79	136.37	61.48
8	4	0.756841	5		4	86.48	78.2	0.34483	87.83	72.07	0.08673	550.17	146.92	66.41
9	5	0.889053	1		5	86.96	79.4	0.31447	92.6	74.81	0.08554	758.28	154.93	82.26

图 11-29　层次分析界面

温馨提示：该程序设置最大可适应于 100 个评价对象，50 个高优指标。用户只需修改图 11-29 中单元格区域 F4:BC4 的权重，以及单元格区域 F5:BC104 范围内指标值数据，便可获得评价对象的排序结果。注意：单元格区域 F5:BC104 范围内指标值应该是高优指标，如果是低优指标，则应进行求倒数转换。

第 12 章

预 测 分 析

预测的方法很多，第 10 章介绍的回归模型就是其中的预测方法之一。本章将重点介绍自回归法、移动平均法、指数平滑法、季节指数法、Markov 链法、灰色预测模型等预测方法。由于圆形分布资料与季节资料有关，所以圆形分布分析也放在本章介绍。

12.1 自相关与自回归

自相关与自回归是在同一时间序列数据基础上建立的相关与回归，一旦数据确定后，其统计分析方法与第 10 章介绍的相关与回归方法相同。

例 12-1：某地 1987～2001 年某病的发病率如表 12-1 所示，求该序列的自相关系数与自回归模型，并预测该病 2002 年、2003 年的发病率。

表 12-1 某地 1987～2001 年某病的发病率（1/10 万）

年份	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001
发病率	11.2	9.7	10.6	8.3	11.5	8.5	8.9	11.2	16.5	17.5	16.2	18.4	21.6	18.2	17.6

12.1.1 线性时序趋势

为了查看表 12-1 中数据的时序趋势情况，采用“数据点折线图”绘制的图形如图 12-1 所示。

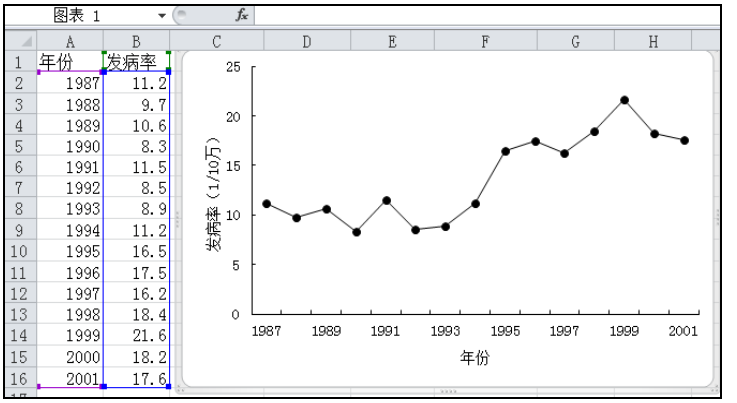


图 12-1 某地 1987～2001 年某病的发病率时序趋势

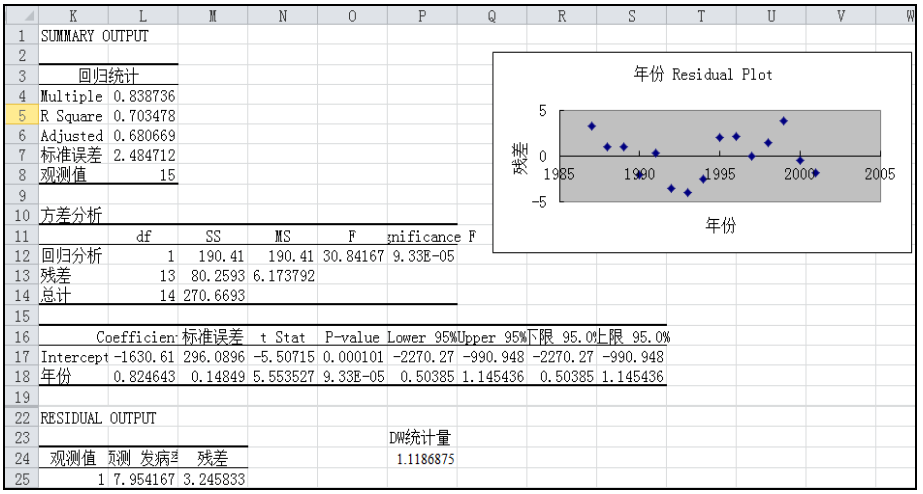
由图 12-1 可见（见光盘文件“12-1 自相关与自回归.xlsx），该数据有一定的线性增长趋势，按第 10 章介绍的方法进行线性回归分析。选定“数据→分析→数据分析→回归”后，单击“确定”按钮，在弹出的“回归”对话框中的“Y 值输入区域”选择$B$1:$B$16，“X 值输入区域”选择$A$1:$A$16，选择“标志”复选框。单击“输出区域”，在其右侧文本框内指定单元格$K$1。在残差选项部分选择残差（residuals）和残差图，然后单击“确定”按钮，获得结果如图 12-2 所示。

由此得到决定系数 R^2=0.70，直线回归模型为：$\widehat{\text{发病率}} = -1630.61 + 0.82\text{年份}$。经方差

分析、回归系数的 t 检验均不拒绝回归模型成立，即发病率与年份之间有一定的线性关系。

将年份代入方程可采用回归方法预测发病率，如将年份 2002 代入回归方程有：$\widehat{发病率}=-1630.61+0.82\times2002=20.3$（1/10 万）；如将年份 2003 代入回归方程有：$\widehat{发病率}=-1630.61+0.82\times2003=21.1$（1/10 万）。

实际发病率与预测发病率之差即为残差（residual）。以年份为横轴，残差为纵轴，可绘制残差图，如图 12-2 的右上侧所示（见光盘文件“12-1 自相关与自回归.xlsx”的“时序趋势”工作表）。由图 12-2 可见，残差散点并不是均匀散在地分布在 0 左右，而是先下降、再上升、再下降的趋势，这说明残差不独立，违背了线性回归残差独立的假定。这种残差趋势是不是由时间序列内部相继序列间（自身）存在一定的相关性造成的呢？下面采用 DW（Durbin-Watson）统计量定量这种相关性。



	K	L	M	N	O	P	Q	R	S
1	SUMMARY OUTPUT								
2									
3	回归统计								
4	Multiple	0.838736							
5	R Square	0.703478							
6	Adjusted	0.680669							
7	标准误差	2.484712							
8	观测值	15							
9									
10	方差分析								
11		df	SS	MS	F	gnificance F			
12	回归分析	1	190.41	190.41	30.84167	9.33E-05			
13	残差	13	80.2593	6.173792					
14	总计	14	270.6693						
15									
16		Coefficien	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
17	Intercept	-1630.61	296.0896	-5.50715	0.000101	-2270.27	-990.948	-2270.27	-990.948
18	年份	0.824643	0.14849	5.553527	9.33E-05	0.50385	1.145436	0.50385	1.145436
19									
22	RESIDUAL OUTPUT								
23						DW统计量			
24	观测值	预测 发病率	残差			1.1186875			
25	1	7.954167	3.245833						

图 12-2　回归分析残差图

12.1.2　DW 统计量

DW 统计量（Dubin-Watson statistics）用于验证回归残差是否独立，是否存在自相关性。计算公式为：

$$\frac{\sum_{t=1+s}^{n}\left(e_t-e_{t-s}\right)^2}{\sum_{t=1}^{n}e_t^2}$$

其中 n 为时序个体数，t 为时序个体编号，e 为残差，s 为自相关差分阶数。DW 统计量的分母实际上就是残差平方和，对于由数据分析工具“回归”获得的残差（见图 12-3 中单元格区域 M25:M39），有

$$3.24583^2 + 0.92119^2 + \cdots + (-1.8992)^2 = 80.25930$$

可采用 Excel 公式“=SUMSQ(M25:M39)”计算。

DW 统计量的分子计算要复杂些，需要考虑自相关差分阶数 s。残差的一阶差分 s=1，对于由数据分析工具“回归”获得的残差（见图 12-3 中单元格区域 M25:M39），有

$$(0.92119 - 3.24583)^2 + (0.99655 - 0.92119)^2 + \cdots + (-1.8992 + 0.4745)^2 = 89.78507$$

可采用 Excel 公式“=SUMXMY2(M26:M39,M25:M38)”计算，在 SUMXMY2 函数中 XMY2 表示 X 减（minus）Y 的差值之平方。

如果将 DW 统计量存放在单元格 P24 中，在单元格 P24 中键入公式“=SUMXMY2(M26:M39,M25:M38)/SUMSQ(M25:M39)”，单击“确定”按钮可得到一阶差分的 DW 统计量值=1.1187。

DW 统计量的取值范围为 0～4。如果值=2，说明不存在自相关性，值越接近于 0 说明越有较强的正自相关；值越接近于 4 说明越有较强的负自相关。本例 DW 统计量为 1.1187，说明残差有一定的正自相关性。

RESIDUAL OUTPUT

			DW统计量
观测值	预测 发病率	残差	1.1186875
1	7.954167	3.245833	
2	8.77881	0.92119	
3	9.603452	0.996548	
4	10.4281	-2.1281	
5	11.25274	0.247262	
6	12.07738	-3.57738	
7	12.90202	-4.00202	
8	13.72667	-2.52667	
9	14.55131	1.94869	
10	15.37595	2.124048	
11	16.2006	-0.0006	
12	17.02524	1.374762	
13	17.84988	3.750119	
14	18.67452	-0.47452	
15	19.49917	-1.89917	

图 12-3 残差及 DW 统计量

12.1.3 自相关

对于同一个时间序列（如表 12-1 所示），如果某个时期（如 1989～2001 年）的数值与该序列向后推移（即滞后，lag）若干时期的数值间有密切联系，就说明该序列存在自相关性（autocorrelation）。令发病率原数据为 Y_t，滞后长度为 1 年的发病率数据为 Y_{t-1}，…，滞后长度为 5 年的发病率数据为 Y_{t-5}，…。则 Y_t、Y_{t-1}、Y_{t-5} 三个变量的数据可表示为：

Y_t	11.2	9.7	10.6	8.3	11.5	8.5	8.9	11.2	16.5	17.5	16.2	18.4	21.6	18.2	17.6
Y_{t-1}		11.2	9.7	10.6	8.3	11.5	8.5	8.9	11.2	16.5	17.5	16.2	18.4	21.6	18.2
Y_{t-5}						11.2	9.7	10.6	8.3	11.5	8.5	8.9	11.2	16.5	17.5

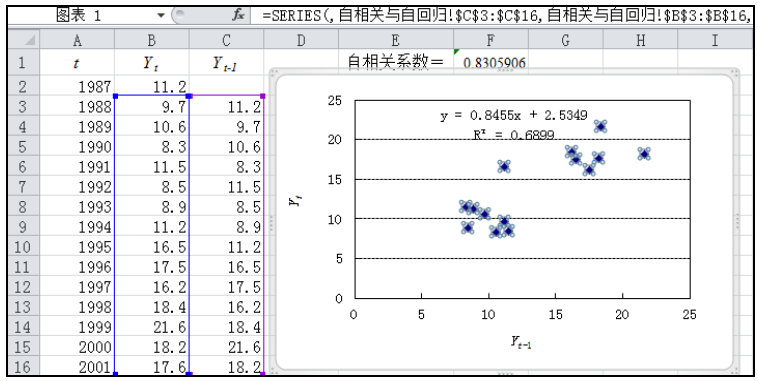
注意：分析时应将其中一个变量无对应关系的数值删除，保持两个变量一对一的关系。例如由 Y_{t-1} 与 Y_t 变量可见，Y_t 的第一个值 11.2 与 Y_{t-1} 的最后一个值 17.6 无对应数值，因此分析时应删除这两端的值。

下面以 Y_{t-1} 与 Y_t 变量为例说明自相关系数的计算。

① 打开光盘文件“12-1 自相关与自回归.xlsx”，在工作表名称上单击鼠标右键选择“插入→工作表”命令，创建一张新的工作表“自相关与自回归”。单元格区域 A2:B16 的数据由“时序趋势”工作表复制、粘贴而得，修改单元格 A1、B1 名称为 t、Y_t，再复制 B2:B15 数据粘贴到 C3:C16，以便创建 Y_{t-1}，在单元格 C1 中键入该列名称 Y_{t-1}。

② 计算 Y_t 与 Y_{t-1} 间的自相关系数。在单元格 E1 中键入“相关系数=”，单击单元格 F1，在编辑栏输入公式“=CORREL(B3:B16,C3:C16)”，单击“确定”按钮，得到自相关系数为 0.8306。

③ 为了直观反映 Y_t 与 Y_{t-1} 间的自相关关系，单击“插入→图表→散点图”命令，在弹出的图表类型界面选择 *XY* 散点图，选取子图表类型中的“散点图”（第一个子图）。以单元格区域 C3:C16 为 *X* 轴，单元格区域 B3:B16 为 *Y* 轴创建散点图，略作适当编辑修饰，可得到图 12-4。



	A	B	C
1	t	Y_t	Y_{t-1}
2	1987	11.2	
3	1988	9.7	11.2
4	1989	10.6	9.7
5	1990	8.3	10.6
6	1991	11.5	8.3
7	1992	8.5	11.5
8	1993	8.9	8.5
9	1994	11.2	8.9
10	1995	16.5	11.2
11	1996	17.5	16.5
12	1997	16.2	17.5
13	1998	18.4	16.2
14	1999	21.6	18.4
15	2000	18.2	21.6
16	2001	17.6	18.2

图 12-4　自相关系数与散点图

由图 12-4 的散点图及自相关系数结果可见：原数据发病率与上一年度发病率间有较强的相关性，也就是说，当某年度的发病率很低时，很可能其下一个年度的发病率也很低；当某年度的发病率较高时，很可能其下一个年度的发病率也较高。

12.1.4　自回归

如果以 Y_t 作为应变量，其余变量（如 Y_{t-1}、Y_{t-5}）作为自变量，来建立回归模型，这种回归模型称为自回归（autoregression，AR）模型。由此可建立一元或多重自回归模型。

引入的最高阶自回归阶数为 1，则为 AR（1）模型；引入的最高阶自回归阶数为 P，则为 AR（P）模型。下面利用图 12-5 的散点图建立 AR（1）模型，即以 Y_{t-1} 为自变量，Y_t 为应变量建立模型。具体操作步骤如下：

①将鼠标移至散点图中的任何一个数据点，光标下方会弹出一个文本框，说明该点的坐标值。

②此时，单击鼠标右键，在弹出的快捷功能区中选择“添加趋势线”。在弹出的“设置趋势线格式”对话框中的“趋势预测/回归分析类型”选择“线性”（默认选项）；在对话框下方的单选框中选择“显示公式（E）”和“显示 R 的平方值（R）”，单击“确定”按钮，此时获得直线回归方程为：$\hat{Y}_t = 2.5349 + 0.8455Y_{t-1}$，决定系数 $R^2 = 0.6899$（如图 12-5 所示，也可见光盘文件“12-1 自相关与自回归.xlsx”中的“自相关与自回归”工作表）。

给定一个 Y_{t-1} 的值，将此值代入自回归方程可预测未来一年的发病率。如预测 2002 年该病的发病率为：$\hat{Y}_{2002} = 2.5349 + 0.8455Y_{2001} = 2.5349 + 0.8455 \times 17.6 = 17.4$。基于 2002 年的发病率预测值 17.4，可预测 2003 年的发病率为：$\hat{Y}_{2003} = 2.5349 + 0.8455Y_{2002} = 2.5349 + 0.8455 \times 17.4 = 17.2$。

如果在单元格 K3 中键入公式“=2.5349+0.8455*C3”，将该公式拖放至 K18，在单元格 C17、C18 中分别键入公式“=B16”、“=K17”，可完成上述预测发病率的计算（如图 12-5 所示）。

为了比较原数据发病率（单元格 B3:B16）与预测发病率（单元格 K3:K16）随年份改变的变化趋势，可采用“带数据标记的折线图”将它们绘制在一张图上，如图 12-5 所示。

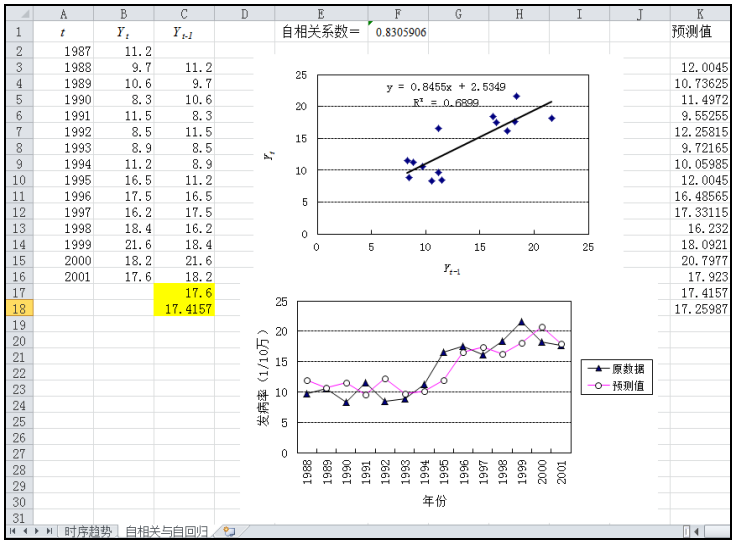


图 12-5 自相关与自回归

AR（2）模型由一个应变量 Y_t，两个自变量 Y_{t-1}、Y_{t-2} 或一个自变量 Y_{t-2} 组成，是否需要拟合该模型，需考察各偏相关系数与 0 有无统计学差异，如果 Y_{t-2} 对应的偏相关系数经检验无统计学意义，则无须拟合 AR（2）模型。是否需要拟合 AR（P）模型，可依此类推。

12.2 移动平均法

移动平均（Moving Average,MA）是一种简单的平滑预测技术。可采用 Excel 的数据分析工具获得。

例 12-2：用表 12-1 中的数据分别求出 3 年、4 年的移动平均，对 2002 年该病发病率作出预测，并在同一个图中比较移动平均与原数据的发病率曲线。

解：求移动平均值按如下步骤进行。

①将表 12-1 中的数据输入到单元格区域 A1:B16 中，在单元格 C1、D1、E1、F1 中分别键入 3 年移动平均、3 年标准误、4 年移动平均、4 年标准误。

②选中功能区“数据→分析→数据分析→移动平均”后，单击“确定”按钮，弹出“移动平均”对话框（见图 12-6）。

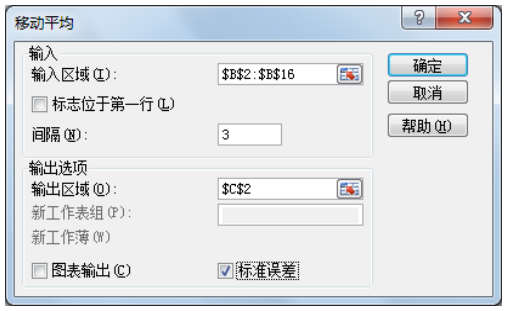


图 12-6 “移动平均”对话框

③“输入区域”选择 B2:B16，在“间隔”右侧文本框中键入 3，“输出区域”选择 C2，选择“标准误差”复选框后单击“确定”按钮。

④按同样方法，且修改“间隔”右侧为“4”，“输出区域”为单元格“E2”或“F2”，可产生 4 年的移动平均和相应的标准误，如图 12-7 所示（可见光盘文件“12-2 移动平均.xlsx”）。

“移动平均”分析工具将计算公式也放到了工作表中，如单元格 C4 中的公式为“=AVERAGE(B2:B4)”，这表明该单元格是由 $(11.2+9.7+10.6)/3=10.5$ 计算得出的；单元格 C5 中的公式为“=AVERAGE(B3:B5)”，由 $(9.7+10.6+8.3)/3=9.5$ 计算得出。每个移动平均计算时使用了 3 个值，即当前行的发病率值及其上面的两个发病率值。

	A	B	C	D	E	F
1	年份	发病率	3年移动平均	3年标准误	4年移动平均	4年标准误
2	1987	11.2				
3	1988	9.7				
4	1989	10.6	10.5			
5	1990	8.3	9.533333333		9.95	
6	1991	11.5	10.13333333	1.06440733	10.025	
7	1992	8.5	9.433333333	1.19163753	9.725	
8	1993	8.9	9.633333333	1.04509436	9.3	1.28050283
9	1994	11.2	9.533333333	1.18133634	10.025	1.14202397
10	1995	16.5	12.2	2.69601901	11.275	2.75417297
11	1996	17.5	15.06666667	3.01047554	13.525	3.34072878
12	1997	16.2	16.73333333	2.8691204	15.35	3.36170994
13	1998	18.4	17.36666667	1.55706276	17.15	3.3684659
14	1999	21.6	18.73333333	1.78605711	18.425	2.65359426
15	2000	18.2	19.4	1.89081601	18.6	1.76957799
16	2001	17.6	19.13333333	2.0007406	18.95	1.84564521

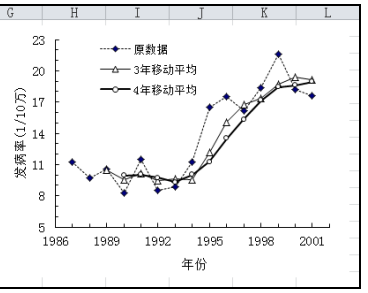


图 12-7 移动平均及其标准误

计算 3 年标准误的单元格 D6 的公式为“=SQRT(SUMXMY2(B4:B6,C4:C6)/3)”。公式中的 SQRT(•)为平方根函数，SUMXMY2(B4:B6,C4:C6)表示三个单元格 B4、B5、B6 的实际值分别与三个单元格 C4、C5、C6 的移动平均值之差的平方和，单元格 D6 实际上就是：

$$\sqrt{[(10.6-10.5)^2+(8.3-9.5)^2+(11.5-10.1)^2]/3}=1.06$$

采用移动平均的最后一个数可预测下一时期的值，可采用标准误衡量预测误差的大小。本例 3 年移动平均预测 2002 年某病发病率为 19.1，标准误为 2.0；4 年移动平均预测 2002 年某病发病率为 19.0，标准误为 1.8，后者预测误差较小。

为了比较原数据发病率（单元格 B2:B16）与 3 年移动平均预测发病率（单元格 C4:C16）和 4 年移动平均预测发病率（单元格 E5:E16）随年份改变的变化趋势，可采用折线散点图，将它们绘制在一张图上，如图 12-7 右侧所示。

由图可见：由于时间序列的原数据受周期变动和随机波动等因素的影响，起伏较大，不易显示出事件的发展趋势；而使用移动平均后，由于移动平均可以平滑数据，消除周期变动和不规则随机波动的影响，使长期趋势显示出来，其中 4 年的移动平均折线比 3 年的移动平均折线更平稳。

计算移动平均时，每次究竟用几项数据来计算，应根据时间序列的总长度和变动周期决定。如果序列总长度较长，周期也长，则可多用几项数据计算移动平均值。一般情况下，季度资料可采用 4 项数据，月份资料可采用 12 项数据计算移动平均。

12.3 指数平滑法

以上所述的移动平均法，实质上是求解当前及以前几项历史值的平滑值。而指数平滑（exponential smoothing）法理论上不是用部分历史值，而是用实际时间序列的所有过去历史值来计算平滑值。

假定时间序列全部数据为 x_t、x_{t-1}、x_{t-2}、…，各期数据的平滑常数为 α，计算 $\hat{x}_{t+1}$ 可采

用下列公式：

$$\begin{aligned}\hat{x}_{t+1} &= \alpha x_t + \alpha(1-\alpha)x_{t-1} + \alpha(1-\alpha)^2 x_{t-2} + \cdots \\ &= \alpha x_t + (1-\alpha)\left[\alpha x_{t-1} + \alpha(1-\alpha)x_{t-2} + \cdots\right] \\ &= \alpha x_t + (1-\alpha)\hat{x}_t\end{aligned} \tag{12-1}$$

也就是：

t+1 时期的预测值=平滑常数×t 时期的实际值+（1–平滑常数）×t 时期的预测值

其中平滑常数 α 的取值范围为 0～1。Excel 称 $(1-\alpha)$ 为阻尼系数（damping coefficient），如果平滑常数 α =0.2，那么阻尼系数 $(1-\alpha)$ =0.8。

12.3.1 数据分析工具——指数平滑

下面采用实例说明数据分析工具——指数平滑的使用方法。

例 12-3：以表 12-1 中数据为例，利用 Excel 指数平滑数据分析工具预测 2002 年某病发病率。

解：令阻尼系数为 0.3（此时平滑常数 α 为 0.7），求指数平滑值按如下步骤进行。

① 将图 12-1 的单元格区域 A1:B16 的数据复制、粘贴到新的工作表中，在单元格 C1、D1 中分别键入文本“预测”、“标准误”。

② 选中“数据→分析→数据分析→指数平滑”后，单击“确定”按钮。

③ 在弹出的“指数平滑”对话框中，“输入区域”选择“B1:B16”；因为指定的输入区域包含标志行，所以选中“标志”复选框；在“阻尼系数”右侧文本框中键入 0.3；“输出区域”选择单元格 C2；选择“标准误差”复选框（如图 12-8 所示）。单击“确定”按钮，即获得如图 12-9 右侧所示结果。

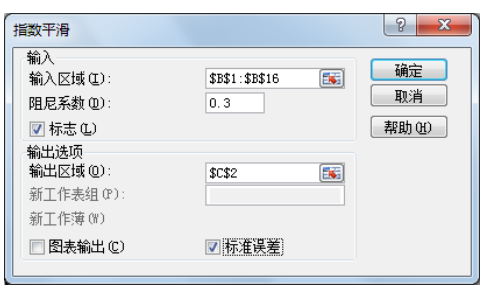
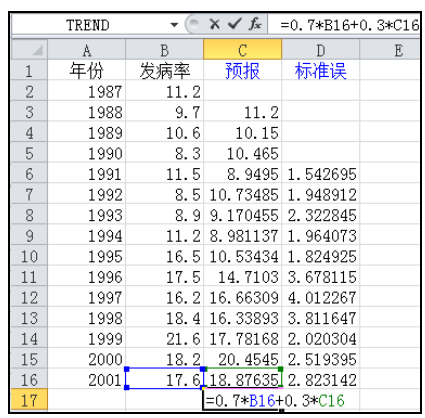


图 12-8 “指数平滑”对话框



	A	B	C	D	E
1	年份	发病率	预报	标准误	
2	1987	11.2			
3	1988	9.7	11.2		
4	1989	10.6	10.15		
5	1990	8.3	10.465		
6	1991	11.5	8.9495	1.542695	
7	1992	8.5	10.73485	1.948912	
8	1993	8.9	9.170455	2.322845	
9	1994	11.2	8.981137	1.964073	
10	1995	16.5	10.53434	1.824925	
11	1996	17.5	14.7103	3.678115	
12	1997	16.2	16.66309	4.012267	
13	1998	18.4	16.33893	3.811647	
14	1999	21.6	17.78168	2.020304	
15	2000	18.2	20.4545	2.519395	
16	2001	17.6	18.87635	2.823142	
17			=0.7*B16+0.3*C16		

图 12-9 指数平滑结果

与“移动平均”分析工具类似，Excel 的“指数平滑”分析工具将计算公式也放在了工作表中。第 C 列计算了指数平滑预测值。单元格 C3 公式为“=B2”，表示将第 1 个时期原数据值作为第 2 个时期的初始预测值；单元格 C4 公式为“=0.7*B3+0.3*C3”，这是采用公式（12-1）预测第 3 个时期的值；单元格 C5 公式为“=0.7*B4+0.3*C4”，……。第 D 列计算预测值相应的标准误，单元格 D6 公式为“=SQRT(SUMXMY2(B3:B5,C3:C5)/3)”，与“移动平均”分析工具的意义相同。单元格 D6 实际上就是：

$$\sqrt{[(9.7-11.2)^2+(10.6-10.2)^2+(8.3-10.5)^2]/3}=1.5$$

单元格 D7 公式为“=SQRT(SUMXMY2(B4:B6,C4:C6)/3)”，……。

为了获得 2002 年的指数平滑预测值，只需先选取单元格区域 C16:D16，将鼠标移向选取单元格右下角的填充柄处，向下拖放至 D17，此时获得预测值为 17.9829，标准误为 2.6640，如图 12-9 所示（可见光盘文件“12-3 指数平滑.xlsx”）。

12.3.2 指数平滑的初始值

以上计算指数平滑预测值采用第 1 个时期的实际值作为第 2 个时期的初始预测值。如果恰好第 1 个时期的实际值变异较大，那么势必对整个指数平滑预测产生较大影响，为此有人建议采用前若干个时期的平均值作为初始预测值。

以下指定前 6 个时期的平均值作为指数平滑的初始值。计算指数平滑预测值的操作如下：

①将图 12-9 复制到新的工作表中，在新工作表中的第 2 行行号上单击鼠标右键，从快捷功能区中选择“插入”，获得一行空白单元格。在单元格 B2 中键入公式“=AVERAGE(B3:B8)”，单击“确定”按钮，该单元格显示数值为 9.966667。将此单元格调整到一位小数，此时显示数值为 10.0。

②选中功能区“数据→分析→数据分析→指数平滑”命令后，单击“确定”按钮。在弹出的“指数平滑”对话框中，输入区域选择 B1:B17；选中“标志”复选框；在“阻尼系数”右侧文本框中键入 0.8（此时平滑常数 α 为 0.2）；输出区域选择单元格 C2；选择“标准误差”。单击“确定”后，弹出是否覆盖已有数据提示框，继续单击“确定”按钮，在单元格区域 C3:D18 产生了新的数据，将这些数据的小数位数调整到 1 位，此时在单元格 C3 中的初始值就是前 6 个时期发病率的平均值，如图 12-10 所示（见光盘文件“12-4 指数平滑 2.xlsx”）。

	A	B	C	D
1	年份	发病率	预报	标准误
2		10.0		
3	1987	11.2	10.0	
4	1988	9.7	10.2	
5	1989	10.6	10.1	
6	1990	8.3	10.2	0.8
7	1991	11.5	9.8	1.2
8	1992	8.5	10.2	1.5
9	1993	8.9	9.8	1.8
10	1994	11.2	9.6	1.5
11	1995	16.5	10.0	1.4
12	1996	17.5	11.3	3.9
13	1997	16.2	12.5	5.3
14	1998	18.4	13.2	5.6
15	1999	21.6	14.3	5.1
16	2000	18.2	15.7	5.6
17	2001	17.6	16.2	5.4
18			16.5	4.5

图 12-10 改变指数平滑的初始值

为了获得 2002 年的指数平滑预测值，只需先选取单元格区域 C17:D17，将鼠标移向选

取单元格右下角的填充柄处，向下拖放至 D18。此时获得预测值为 16.5，标准误为 4.5。

12.3.3 求解最佳平滑常数

图 12-9 与图 12-10 所给出结果的平滑常数α分别为 0.7 和 0.2，其阻尼系数分别为 0.3 和 0.8。这些值都是主观给出的，那么采用什么平滑常数最佳呢？其评判标准是使实际值与预测值间的误差最小。下面以实际值与预测值间差值的均数的平方加上其标准差的平方，然后开方，来衡量这种误差（记为 RMSPE）的大小。用公式可表示为：

$$\text{RMSPE} = \sqrt{\text{差值的均数}^2 + \text{差值的标准差}^2} \qquad (12\text{-}2)$$

求解最佳平滑常数的步骤如下：

①将图 12-1 中的数据复制到一张新的工作表的单元格区域 A1:B16。在单元格 B1、C1、D1、F1、F5、F6、F8 处按图 12-11 所示（见光盘文件“12-5 最佳平滑常数.xlsx”）键入文字。

	A	B	C	D	E	F	G
1	年份	实际发病率	预报发病率	差值		平滑常数	1
2	1987	11.2	=AVERAGE(B2:B7)				
3	1988	9.7	=$G$1*B2+(1-$G$1)*C2				
4	1989	10.6	=$G$1*B3+(1-$G$1)*C3				
5	1990	8.3	=$G$1*B4+(1-$G$1)*C4			差值的均数	=AVERAGE(D8:D16)
6	1991	11.5	=$G$1*B5+(1-$G$1)*C5			差值的标准差	=STDEV(D8:D16)
7	1992	8.5	=$G$1*B6+(1-$G$1)*C6				
8	1993	8.9	=$G$1*B7+(1-$G$1)*C7	=B8-C8		RMSPE	=SQRT(G5^2+G6^2)
9	1994	11.2	=$G$1*B8+(1-$G$1)*C8	=B9-C9			
10	1995	16.5	=$G$1*B9+(1-$G$1)*C9	=B10-C10			
11	1996	17.5	=$G$1*B10+(1-$G$1)*C10	=B11-C11			
12	1997	16.2	=$G$1*B11+(1-$G$1)*C11	=B12-C12			
13	1998	18.4	=$G$1*B12+(1-$G$1)*C12	=B13-C13			
14	1999	21.6	=$G$1*B13+(1-$G$1)*C13	=B14-C14			
15	2000	18.2	=$G$1*B14+(1-$G$1)*C14	=B15-C15			
16	2001	17.6	=$G$1*B15+(1-$G$1)*C15	=B16-C16			
17			=$G$1*B16+(1-$G$1)*C16				

图 12-11 求解最佳平滑常数所需键入的原数据与公式

②在单元格 C2 中键入公式“=AVERAGE(B2:B7)”作为预测值的初始值，在单元格 C3 中键入公式“=$G$1*B2+(1–$G$1)*C2”求 1988 年预测发病率，为了求其他年份的预测发病率，将此公式向下拖放至单元格 C16。其中单元格 G1 为平滑常数值所在的位置。

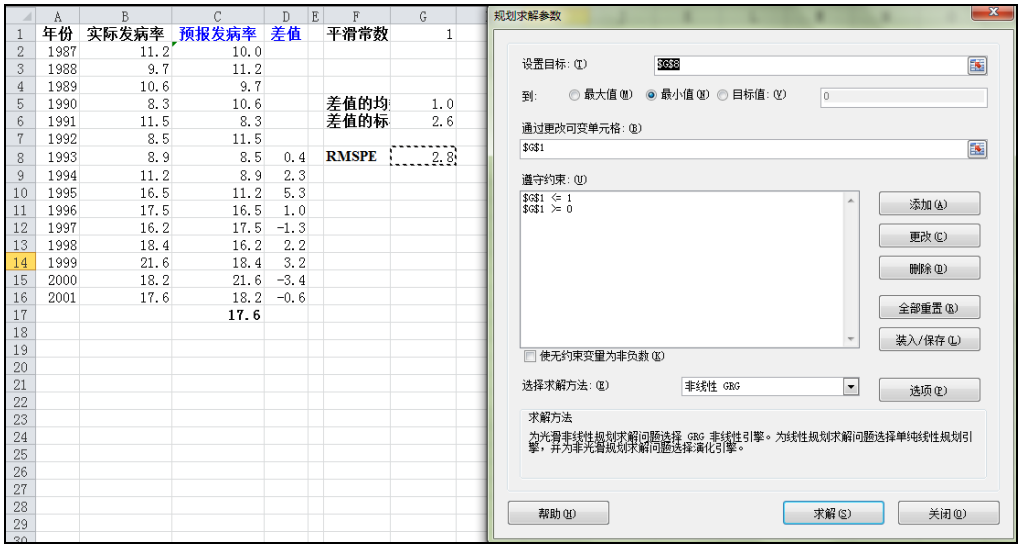
③因为初始值是前面 6 个年份的发病率平均值，所以从第 7 个年份开始计算差值。在单元格 D8 中键入公式“=B8–C8”，并将此公式向下拖放填充至单元格 D16。

④在单元格 G5 中键入公式“=AVERAGE(D8:D16)”计算差值的均数，在单元格 G6 中键入公式“=STDEV(D8:D16)”计算差值的标准差；在单元格 G8 中键入公式“=SQRT(G5^2+G6^2)”计算误差 RMSPE。

要使误差 RMSPE 最小来求解平滑常数，需要用到在约束条件下进行规划求解求极值的问题（参见第 11 章）。具体操作如下：

单击“数据→分析→数据分析→规划求解”命令，按照图 12-12 右侧的“规划求解参

数”对话框填写参数。其中约束条件实际上是限制平滑常数的取值范围在 0~1。单击“求解”按钮，将弹出“规划求解结果”对话框（如图 12-12 所示）。单击“确定”按钮，得到平滑常数的最佳值为 1，即阻尼系数为 0 时，可获得较好的指数平滑预测结果。



	A	B	C	D	E	F	G
1	年份	实际发病率	预报发病率	差值		平滑常数	1
2	1987	11.2	10.0				
3	1988	9.7	11.2				
4	1989	10.6	9.7				
5	1990	8.3	10.6			差值的均	1.0
6	1991	11.5	8.3			差值的标	2.6
7	1992	8.5	11.5				
8	1993	8.9	8.5	0.4		RMSPE	2.8
9	1994	11.2	8.9	2.3			
10	1995	16.5	11.2	5.3			
11	1996	17.5	16.5	1.0			
12	1997	16.2	17.5	-1.3			
13	1998	18.4	16.2	2.2			
14	1999	21.6	18.4	3.2			
15	2000	18.2	21.6	-3.4			
16	2001	17.6	18.2	-0.6			
17			17.6				

图 12-12 “规划求解参数”对话框

此时，将单元格 C16 公式复制到 C17，在单元格 C17 中显示数值为 17.6，在最佳平滑常数 $\alpha=1$ 时，此值就是 2002 年的发病率预测结果。

12.4 季节预测法

按月或季搜集到的时间序列资料通常包括以下变动因素：

①长期趋势（trend）：相当长时间内变动的总方向。

②季节趋势（season）：由季节（如季度、月份）变换引起的数量变化。

③随机趋势（random）：由突发或偶然因素引起的数量变化。

例 12-4：某医院 1997～2001 年各季度的门诊人数，如图 12-13 所示的单元格区域 A1:D21（见光盘文件“12-6 季节门诊人次.xlsx”），试预测 2002 年各季度的门诊人数。

图中第 C 列是这样产生的：完成第 A、B 列数据的输入后，在单元格 C2 处键入公式“=A2&" "&B2”，然后将此公式向下拖放至单元格 C21。以第 C 列为横轴、门诊人数为纵轴，采用“数据点折线图”可绘制如图 12-13 右侧所示图形，由该折线图可见该数据既有向上的长期趋势，又有随季度变化的季节趋势（第三季度最低，第一、四季度较高）。

	A	B	C	D
1	年份	季度	年份季度	门诊人数
2	1997	一	1997 一	2999
3		二	二	3138
4		三	三	2522
5		四	四	2899
6	1998	一	1998 一	3067
7		二	二	2590
8		三	三	2478
9		四	四	3076
10	1999	一	1999 一	3372
11		二	二	2894
12		三	三	2858
13		四	四	3388
14	2000	一	2000 一	3895
15		二	二	3282
16		三	三	3338
17		四	四	3780
18	2001	一	2001 一	4217
19		二	二	3603
20		三	三	3581
21		四	四	3986

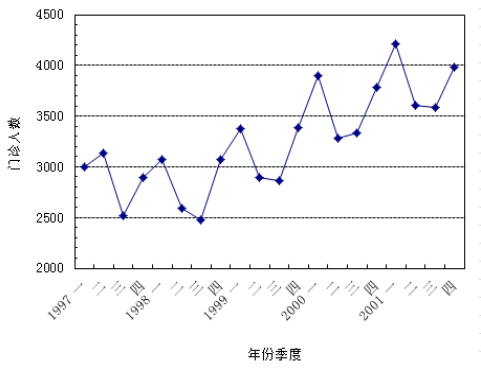


图 12-13　某医院 1997～2001 年各季度的门诊人数

下面采用哑变量多重回归、自回归模型和季节指数法三种方法完成对 2002 年各季度的预测。

12.4.1　哑变量多重回归

将图 12-13 中的单元格区域 A1:D21 复制到一张新的工作表中，按图 12-14（见光盘文件“12-7 哑变量回归.xlsx”中的“回归”工作表）建立哑变量数据。用变量 *time* 反映时间序列数据的长期趋势；用季节哑变量 *s*1、*s*2、*s*3、*s*4（分别对应于季度一、二、三、四，见单元格区域 F1:I21。如季度为二，则 *s*1、*s*2、*s*3、*s*4 分别为 0、1、0、0；季度为四，则 *s*1、*s*2、*s*3、*s*4 分别为 0、0、0、1）反映数据的季节趋势，将分类变量（如季度）哑变量化后，一般只将比分类个数少 1 个哑变量的其他变量放入模型中，如对于季度资料只选 3 个哑变量；对于月份资料只选 11 个哑变量。

	A	B	C	D	E	F	G	H	I
1	年份	季度	年份季度	门诊人数	time	s1	s2	s3	s4
2	1997	一	1997 一	2999	1	1	0	0	0
3		二	二	3138	2	0	1	0	0
4		三	三	2522	3	0	0	1	0
5		四	四	2899	4	0	0	0	1
6	1998	一	1998 一	3067	5	1	0	0	0
7		二	二	2590	6	0	1	0	0
8		三	三	2478	7	0	0	1	0
9		四	四	3076	8	0	0	0	1
10	1999	一	1999 一	3372	9	1	0	0	0
11		二	二	2894	10	0	1	0	0
12		三	三	2858	11	0	0	1	0
13		四	四	3388	12	0	0	0	1
14	2000	一	2000 一	3895	13	1	0	0	0
15		二	二	3282	14	0	1	0	0
16		三	三	3338	15	0	0	1	0
17		四	四	3780	16	0	0	0	1
18	2001	一	2001 一	4217	17	1	0	0	0
19		二	二	3603	18	0	1	0	0
20		三	三	3581	19	0	0	1	0
21		四	四	3986	20	0	0	0	1

图 12-14　哑变量数据的建立

选择“数据→分析→数据分析→回归”命令后，单击“确定”按钮。在弹出的“回归”对话框中，*Y* 值输入区域是 D1:D21，*X* 值输入区域是 E1:H21（只包括变量 *time*、*s*1、*s*2、*s*3），复选“标志”框，在“输出区域”右侧文本框中指定单元格为 K1，选取残差部分的所有复选框。单击“确定”按钮，获得如图 12-15（见光盘文件“12-7 哑变量回归.xlsx”中的“回归”工作表）所示的结果。

K	L	M	N	O	P	Q	R	S
SUMMARY OUTPUT								
回归统计								
Multiple	0.92787							
R Square	0.860942							
Adjusted	0.82386							
标准误差	205.4924							
观测值	20							
方差分析								
	df	SS	MS	F	gnificance F			
回归分析	4	3921579	980394.9	23.21717	2.8E-06			
残差	15	633407.1	42227.14					
总计	19	4554987						
	Coefficien	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
Intercept	2620.15	133.9647	19.55851	4.36E-12	2334.611	2905.689	2334.611	2905.689
time	67.1375	8.122802	8.265313	5.75E-07	49.82415	84.45085	49.82415	84.45085
s1	285.6125	132.2296	2.159974	0.04737	3.771541	567.4535	3.771541	567.4535
s2	-190.125	130.9762	-1.4516	0.167207	-469.294	89.04442	-469.294	89.04442
s3	-403.263	130.2184	-3.09682	0.007365	-680.817	-125.708	-680.817	-125.708

图 12-15 哑变量多重回归分析部分结果

本例决定系数（R Square）为 0.8609，调整决定系数（Adjusted R Square）为 0.8238，说明模型拟合较好，剩余标准差为 205.49，（与平均门诊人数比较）可粗略反映模型预测的误差大小，方差分析 *F* 值对应的 *P* 值远小于 0.001，也表明该回归模型成立。其回归方程可写为：

$$\hat{Y} = 2620.15 + 67.14time + 285.61s1 - 190.13s2 - 403.26s3 \qquad (12\text{-}3)$$

变量 *time* 的偏回归系数为 67.14，说明在扣除了四季度影响后，1997～2001 年该医院平均每季度的门诊量增加了 67 人。回归方程中没有第四季度的哑变量 *s*4，实际上第四季度的季节性影响已经包括在常数项（Intercept）2620.15 中了。*s*1 的系数为 285.61，说明第一季度的门诊人数比第四季度平均多 286 人，*s*2、*s*3 的系数分别为−190.13、−403.26，说明第二季度、第三季度的门诊人数比第四季度分别平均少了 190、403 人。

经 *t* 检验，在检验水准 $\alpha = 0.05$ 时，变量 *time*、*s*1、*s*3 均有统计学意义，只是 *s*2 的 *t* 检验 *P* 值为 0.1672 大于 0.05。

利用公式（12-3），可预测 2002 年四个季度的门诊人数。选取单元格区域 A18:C21，将此复制、粘贴到单元格 A22，并将单元格 A22 中的 2001 改为 2002；选取单元格区域 E20:E21，将此向下拖放填充至单元格 E25；选取单元格区域 F18:I21，将此复制、粘贴到单元格 F22。根据公式（12-3），在单元格 D22 中键入公式：

“=$L$17+$L$18*E22+$L$19*F22+$L$20*G22+$L$21*H22”

可预测 2002 年第一季度该医院的门诊人数为 4316 人，将该公式向下拖放填充至单元格 D25，可预测第二、三、四季度门诊人数分别为 3907、3761、4231 人（如图 12-16 所示，可见光盘文件“12-7 哑变量回归.xlsx”中的“预测”工作表）。

使用 TREND 统计函数也可预测 2002 年的门诊人数。具体操作步骤如下：

①将图 12-16 中的单元格区域 A1:I25 复制到一张新的工作表，清除单元格区域 D22:D25 中的内容，并在单元格 J1 处键入“预测”字样。

②选取单元格 J2:J25，单击函数按钮（图标 *fx*），单击“统计”函数分类，并在右侧函数名中选择“TREND”，单击“确定”后将弹出 TREND 函数参数对话框，按图 12-17 分别填写建立模型的应变量、自变量以及预测所需的自变量。为了使回归系数的常数项不为 0，在 Const 右侧文本框中键入 1，此时不能单击“确定”，而应该按 Ctrl+Shift+Enter 组合键，然后便可在单元格区域 J2:J25 产生该医院 1997～2002 年的预测门诊人数，获得的 2002 年四个季度的预测结果与前面相同，如图 12-18 所示（见光盘文件“12-7 哑变量回归.xlsx”中的“趋势函数”工作表）。

	A	B	C	D	E	F	G	H	I
1	年份	季度	年份季度	门诊人数	time	s1	s2	s3	s4
2	1997	一	1997 一	2999	1	1	0	0	0
3		二	二	3138	2	0	1	0	0
4		三	三	2522	3	0	0	1	0
5		四	四	2899	4	0	0	0	1
22	2002	一	2002 一	4315.65	21	1	0	0	0
23		二	二	3907.05	22	0	1	0	0
24		三	三	3761.05	23	0	0	1	0
25		四	四	4231.45	24	0	0	0	1

图 12-16　采用哑变量多重回归模型公式预测

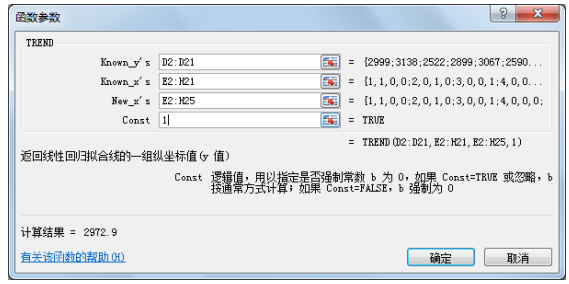


图 12-17　TREND 函数参数对话框

J25　{=TREND(D2:D21,E2:H21,E2:H25,1)}

	A	B	C	D	E	F	G	H	I	J
1	年份	季度	年份季度	门诊人数	time	s1	s2	s3	s4	预测
2	1997	一	1997 一	2999	1	1	0	0	0	2972.9
3		二	二	3138	2	0	1	0	0	2564.3
4		三	三	2522	3	0	0	1	0	2418.3
5		四	四	2899	4	0	0	0	1	2888.7
6	1998	一	1998 一	3067	5	1	0	0	0	3241.45
7		二	二	2590	6	0	1	0	0	2832.85
8		三	三	2478	7	0	0	1	0	2686.85
9		四	四	3076	8	0	0	0	1	3157.25
10	1999	一	1999 一	3372	9	1	0	0	0	3510
11		二	二	2894	10	0	1	0	0	3101.4
12		三	三	2858	11	0	0	1	0	2955.4
13		四	四	3388	12	0	0	0	1	3425.8
14	2000	一	2000 一	3895	13	1	0	0	0	3778.55
15		二	二	3282	14	0	1	0	0	3369.95
16		三	三	3338	15	0	0	1	0	3223.95
17		四	四	3780	16	0	0	0	1	3694.35
18	2001	一	2001 一	4217	17	1	0	0	0	4047.1
19		二	二	3603	18	0	1	0	0	3638.5
20		三	三	3581	19	0	0	1	0	3492.5
21		四	四	3986	20	0	0	0	1	3962.9
22	2002	一	2002 一		21	1	0	0	0	4315.65
23		二	二		22	0	1	0	0	3907.05
24		三	三		23	0	0	1	0	3761.05
25		四	四		24	0	0	0	1	4231.45

图 12-18　TREND 统计函数预测结果

12.4.2 AR（4）模型

为了对季节性资料建立自回归模型，首先需要计算自相关系数函数（Autocorrelation Coefficients Function，ACF），以便初步确定自回归模型的阶数。按下列步骤建立 ACF。

①在工作表名称上单击鼠标右键，创建一个新的工作表“自回归与 ACF”。将图 12-14 中单元格区域 A1:D21 数据复制到新的工作表中（参见光盘文件“12-8 AR 模型.xlsx”）。

②在单元格 E1、G1、H1 处分别键入文字 *z* 值、滞后、ACF；在单元格 G2:G9 中输入数字 1～8。其中 *z* 值就是标准正态离差值。

③单击“公式→定义的名称→名称管理器”命令，在弹出的“名称管理器”对话框中单击“新建”按钮，弹出“新建名称”对话框，在“名称”框写“门诊人数”，在“引用位置”选取单元格区域 D2:D21，建立“门诊人数”的数据范围命名；以类似方式可将单元格区域 E2:E21 命名为“z 值”。

④在单元格 E2 中键入公式“=(D2−AVERAGE(门诊人数))/STDEV(门诊人数)”，以便获得标准正态离差值 *z*。单击单元格 E2，将此公式向下拖放填充至单元格 E21，调整小数位数至 3 位。

⑤在单元格 H2 中键入公式“=SUMPRODUCT(OFFSET(*z* 值,G2,0,20-G2),OFFSET(*z* 值,0,0,20-G2))/19”，并将此公式向下拖放至单元格 H9，调整小数位数至 4 位。

⑥选取单元格区域 H2:H9，单击“图表向导”图标，选取柱形图，可绘制 ACF 图，如图 12-19 右侧所示（见光盘文件“12-8 AR 模型.xlsx”中“自回归与 ACF”工作表）。

	A	B	C	D	E	G	H
1	年份	季度	年份季度	门诊人数	Z值	滞后	ACF
2	1997	一	1997 一	2999	-0.509	1	0.5680
3		二	二	3138	-0.225	2	0.3051
4		三	三	2522	-1.483	3	0.4899
5		四	四	2899	-0.713	4	0.5253
6	1998	一	1998 一	3067	-0.370	5	0.1405
7		二	二	2590	-1.344	6	-0.1199
8		三	三	2478	-1.573	7	0.0084
9		四	四	3076	-0.352	8	-0.0197
10	1999	一	1999 一	3372	0.253		
11		二	二	2894	-0.723		
12		三	三	2858	-0.797		
13		四	四	3388	0.286		
14	2000	一	2000 一	3895	1.321		
15		二	二	3282	0.069		
16		三	三	3338	0.184		
17		四	四	3780	1.086		
18	2001	一	2001 一	4217	1.979		
19		二	二	3603	0.725		
20		三	三	3581	0.680		
21		四	四	3986	1.507		

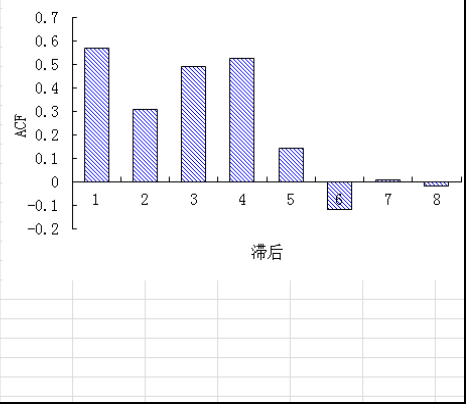


图 12-19 自相关系数函数（ACF）直条图

由 ACF 值及其图形可见，滞后 1、滞后 4、滞后 3 的 ACF 值较高，本例可选择 AR（4）模型（在实际应用中，尚需结合偏自相关函数与逆自相关函数提供的信息，此处简化了具

体模型识别过程）。因为包括滞后 1 和滞后 3 的自变量后，其模型相应偏回归系数在检验水准 0.05 时无统计学意义，所以下面的 AR（4）模型没有包括滞后 1 和滞后 3 作为自变量。具体建模方法如下：

①按图 12-20 所示（可见光盘文件“12-8 AR 模型.xlsx”）在一张新的工作表（AR(4)）中建立自回归数据。

②选择“数据→分析→数据分析”命令，选取“回归”，单击“确定”按钮。在“回归”对话框中，Y 值输入区域是 D6:D21，X 值输入区域是 E6:E21（此为 D 列数据后移 4 个数字），指定输出区域为单元格 G1，单击“确定”按钮，获得如图 12-20 右侧所示结果。

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	年份	季度	年份季度	门诊人数Y_t	门诊人数Y_{t-4}		SUMMARY OUTPUT								
2	1997	一	1997 一	2999											
3		二	二	3138			回归统计								
4		三	三	2522			Multiple	0.860896							
5		四	四	2899			R Square	0.741142							
6	1998	一	1998 一	3067	2999		Adjusted	0.722652							
7		二	二	2590	3138		标准误差	261.662							
8		三	三	2478	2522		观测值	16							
9		四	四	3076	2899										
10	1999	一	1999 一	3372	3067		方差分析								
11		二	二	2894	2590			df	SS	MS	F	nificance F			
12		三	三	2858	2478		回归分析	1	2744410	2744410	40.08369	1.86E-05			
13		四	四	3388	3076		残差	14	958538.2	68467.01					
14	2000	一	2000 一	3895	3372		总计	15	3702948						
15		二	二	3282	2894										
16		三	三	3338	2858			Coefficien	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
17		四	四	3780	3388		Intercept	76.25577	519.2954	0.146845	0.885348	-1037.52	1190.034	-1037.522	1190.034
18	2001	一	2001 一	4217	3895		X Variabl	1.052624	0.166261	6.331168	1.86E-05	0.696031	1.409218	0.6960307	1.409218
19		二	二	3603	3282										
20		三	三	3581	3338										
21		四	四	3986	3780										
22				4515.1729	4217										
23				3868.861513	3603										
24				3845.703777	3581										
25				4272.016661	3986										

图 12-20　自回归模型数据及其回归结果

模型决定系数为 0.74，剩余标准差为 261.66。可见 AR（4）模型比前面建立的哑变量多重回归模型要差。

单元格区域 E22:E25 是原数据 2001 年的门诊人数，在单元格 D22 中键入公式“=$H$17+$H$18*E22”，然后将此单元格公式向下拖放填充至 D25，可得到 2002 年第一、二、三、四季度的预测门诊人数分别为：4515、3869、3846、4272（见光盘文件“12-8AR 模型.xlsx”中的“AR（4）”工作表）。

12.4.3　季节指数法

该方法的基本思想是：原数据包含长期趋势（T）、季节趋势（S）、随机趋势（R）等，而移动平均平滑了原时间序列数据，反映了长期趋势和随机趋势（TR）。如果用原数据（TSR）除以移动平均值（TR），将可以获得反映季节趋势的值，该值称为季节指数（seasonal index）。

（1）校正季节指数

校正季节指数也叫做标准化季节指数，可采用如下步骤获得。

①将图 12-20 中单元格区域 A1:D21 中的数据复制到一张新的工作表中，按图 12-21 在单元格 E1、F1、G1:G6、H1、I1 中输入文本。

②在单元格 E4 中键入公式“=AVERAGE(D2:D5,D3:D6)”，并将该单元格建立的公式拖放填充至单元格 E19，获得居中移动平均值（数字小数位数调整为 0）。

	A	B	C	D	E	F	G	H	I
1	年份	季度	年份季度	门诊人数	居中移动平均	季节指数	季度	平均季节指数	校正季节指数
2	1997	一	1997 一	2999			一	1.1233	1.1289
3		二	二	3138			二	0.9356	0.9403
4		三	三	2522	2898	0.8703	三	0.8903	0.8947
5		四	四	2899	2838	1.0215	四	1.0309	1.0361
6	1998	一	1998 一	3067	2764	1.1096	合计	3.9801	4.0000
7		二	二	2590	2781	0.9314			
8		三	三	2478	2841	0.8723			
9		四	四	3076	2917	1.0545			
10	1999	一	1999 一	3372	3003	1.1231			
11		二	二	2894	3089	0.9369			
12		三	三	2858	3193	0.8950			
13		四	四	3388	3307	1.0244			
14	2000	一	2000 一	3895	3416	1.1403			
15		二	二	3282	3525	0.9311			
16		三	三	3338	3614	0.9236			
17		四	四	3780	3694	1.0232			
18	2001	一	2001 一	4217	3765	1.1201			
19		二	二	3603	3821	0.9429			
20		三	三	3581					
21		四	四	3986					

图 12-21 校正季节指数

③在单元格 F4 中键入公式“=D4/E4”，并将该单元格建立的公式拖放填充至单元格 F19，获得季节指数（参见光盘文件“12-9 季节指数法.xlsx”）。

④在单元格 H2 中键入公式“=AVERAGE(F6,F10,F14,F18)”，复制单元格 H2 并粘贴至单元格 H3 中，可分别获得第一、二季度平均季节指数。在单元格 H4 中键入公式“=AVERAGE(F4,F8,F12,F16)”，复制单元格 H4 并粘贴至单元格 H5 中，可分别获得第三、四季度平均季节指数。选中单元格 H6，在功能区“开始”下的“编辑”命令组，单击自动求和符号，可获得 4 个平均季节指数的和为 3.9801。

⑤在单元格 I2 中键入公式“=4*H2/$H$6”，然后将该公式向下拖放填充至 I5，可获得 4 个季度的校正季节指数。单击自动求和符号，获得 4 个校正季节指数的和为 4。

（2）预测

按如下步骤进行预测。

①复制单元格区域 I2:I5，选择性粘贴数值（消除公式）至单元格 K2。然后选取并复制单元格 K2:K5，分别粘贴到单元格 K6、K10、K14、K18、K22 中。

②选取单元格 L2，输入公式“=D2/K2”，并将该公式向下拖放填充至单元格 L21。此时单元格区域 L2:L21 中的值即为经季节调整后的原数据，即长期趋势数据。

③选中单元格区域 L2:L21，复制后选择性粘贴数值，以便去除该列中的公式。

④在单元格 N1 处键入字母 *t*，单元格区域 N2:N25 产生自然数 1、2、……、24。以便

使 t 作为自变量，单元格区域 L2:L21 中数据作为应变量，产生简单线性回归趋势模型，预测单元格区域 L22:L25 中的值。

⑤选取单元格区域 L22:L25，单击函数 *fx* 图标，选择统计函数 TREND，在 TREND 函数参数对话框中的 Known_y's 后填 L2:L21，Known_x's 后填 N2:N21，New_x's 后填 N22:N25，Const 后填 1，按 Ctrl+Shift+Enter 组合键，得单元格区域 L22:L25 的趋势预测值分别为 3955、4022、4089、4155。

⑥在单元格 M22 中键入公式“=K22*L22”，并将此公式向下拖放填充至 M25，可获得 2002 年第一、二、三、四季度的预测值分别为 4465、3782、3658、4305，如图 12-22 所示（见光盘文件“12-9 季节指数法.xlsx”）。

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	年份	季度	年份季度	门诊人数	居中移动平均	季节指数	季度	平均季节指数	校正季节指数		指数	趋势	预报	t
2	1997	一	1997	一 2999			一	1.1233	1.1289		1.1289	2657		1
3		二		二 3138			二	0.9356	0.9403		0.9403	3337		2
4		三		三 2522	2898	0.8703	三	0.8903	0.8947		0.8947	2819		3
5		四		四 2899	2838	1.0215	四	1.0309	1.0361		1.0361	2798		4
6	1998	一	1998	一 3067	2764	1.1096	合计	3.9801	4.0000		1.1289	2717		5
7		二		二 2590	2781	0.9314					0.9403	2754		6
8		三		三 2478	2841	0.8723					0.8947	2770		7
19		二		二 3603	3821	0.9429					0.9403	3832		18
20		三		三 3581							0.8947	4002		19
21		四		四 3986							1.0361	3847		20
22	2002	一	2002	一							1.1289	3955	4465	21
23		二		二							0.9403	4022	3782	22
24		三		三							0.8947	4089	3658	23
25		四		四							1.0361	4155	4305	24

图 12-22　季节指数预测结果

对于同一资料，采用哑变量多重回归、自回归、季节指数法三种方法获得的 2002 年 4 个季度的预测结果不尽相同（如表 12-2 所示）。具体选用哪种预测方法，需结合实际的具体情况而定；也可以利用基于残差的各种判别准则进行模型筛选（参见有关书籍）。

表 12-2　三种方法预测某医院 2002 年 4 个季度的门诊人数

季　　度	哑变量多重回归	AR（4）模型	季节指数法
一	4316	4515	4465
二	3907	3869	3782
三	3761	3846	3658
四	4231	4272	4305

12.5　Markov 预测法

A. A. Markov 是俄国著名数学家，他于 20 世纪初发现：一些事物的变化与事物的近期状态关系极大，而与事物的远期状态关系很小，甚至可以忽略不计，这种现象叫做“无后效性”。利用这一现象进行预测的方法称 Markov 预测法。

药品销售中有“畅销”、“滞销”等状态，机器运转有“正常”、“故障”等状态。状态

不是固定不变的，一种状态可以向另外一种状态转移。从某种状态出发，下一步转移到其他状态的可能性大小，叫做“状态转移概率”。比如某医疗设备第1年“正常”到第2年仍然“正常”的概率为0.8，则“故障”概率为0.2；第1年出现“故障”到第2年“正常”的概率为0.7，则“故障”概率为0.3（如表12-3所示），这些小数就是状态转移概率。矩阵$\begin{vmatrix}0.8 & 0.2\\ 0.7 & 0.3\end{vmatrix}$称为状态转移概率矩阵，用符号$\boldsymbol{P}$表示。注意：系统各个状态是互斥的，转移概率矩阵中每一行元素之和为1，如0.8+0.2=1。

表12-3 某医疗设备状态转移概率

	正 常	故 障
正常	0.8	0.2
故障	0.7	0.3

一旦确定了转移概率矩阵，就可以根据Markov过程的重要性质：$\boldsymbol{P}_k=\boldsymbol{P}_0\boldsymbol{P}^k$进行预测。其中$\boldsymbol{P}_0$表示事物的初始状态概率，$\boldsymbol{P}$为状态转移概率矩阵，$\boldsymbol{P}_k$为经$k$步转移后的状态概率。

例12-5：某研究者对592例慢性肾病患者随访一年后，得到各类慢性肾病相互转移的例数如表12-4a所示。如在原分类为1的200例慢性肾病患者中，一年后仍有146人为1类，转移为2类的有51人，转移为3类的有3人。表12-4b是由表12-4a产生的，如146/200=0.730，51/200=0.255，3/200=0.015，每行频率相加为1，如0.730+0.255+0.015+0+0=1。

某医院现已获得1、2、3、4、5类慢性肾病病人各有43、28、19、14、2人。如果采用表12-4b作为转移概率矩阵，试预测随访1年、2年、3年后各类慢性肾病病人的数量。

表12-4a 5类慢性肾病相互转移的例数

类别	1	2	3	4	5	合计
1	146	51	3	0	0	200
2	47	75	28	1	0	151
3	2	17	65	17	0	101
4	0	3	9	31	37	80
5	0	0	0	0	60	60

表12-4b 5类慢性肾病相互转移的频率

类别	1	2	3	4	5
1	0.730	0.255	0.015	0.000	0.000
2	0.311	0.497	0.185	0.007	0.000
3	0.020	0.168	0.644	0.168	0.000
4	0.000	0.038	0.113	0.388	0.463
5	0.000	0.000	0.000	0.000	1.000

解：现共有病人（43+28+19+14+2）=106人，病人分类的初始状态$\boldsymbol{P}_0$=(43/106,28/106, 19/106,14/106,2/106)=(0.406, 0.264,0.179, 0.132, 0.019)。

（1）1 年转移后的状态概率

$$\boldsymbol{P}_1=\boldsymbol{P}_0\boldsymbol{P}^1=(0.406,0.264,0.179,0.132,0.019)\begin{vmatrix}0.730 & 0.255 & 0.015 & 0 & 0\\ 0.311 & 0.497 & 0.185 & 0.007 & 0\\ 0.020 & 0.168 & 0.644 & 0.168 & 0\\ 0 & 0.038 & 0.113 & 0.388 & 0.463\\ 0 & 0 & 0 & 0 & 1\end{vmatrix}$$

$$=(0.3819,0.2698,0.1853,0.0831,0.0800)$$

采用 106 乘以 $\boldsymbol{P}_1$ 各概率可预测得 1 年转移后的 1、2、3、4、5 类慢性肾病人数分别为 40、29、20、9、8。Excel 计算的部分公式如图 12-23 所示，其相应计算结果如图 12-24 所示。

	A	B	C	D	E	F	G	H
1		类别	1	2	3	4	5	合计
2		1	146	51	3	0	0	=SUM(C2:G2)
3		2	47	75	28	1	0	=SUM(C3:G3)
4		3	2	17	65	17	0	=SUM(C4:G4)
5		4	0	3	9	31	37	=SUM(C5:G5)
6		5	0	0	0	0	60	=SUM(C6:G6)
7								=SUM(H2:H6)
8		类别	1	2	3	4	5	
9	P	1	=C2/$H2	=D2/$H2	=E2/$H2	=F2/$H2	=G2/$H2	
10		2	=C3/$H3	=D3/$H3	=E3/$H3	=F3/$H3	=G3/$H3	
11		3	=C4/$H4	=D4/$H4	=E4/$H4	=F4/$H4	=G4/$H4	
12		4	=C5/$H5	=D5/$H5	=E5/$H5	=F5/$H5	=G5/$H5	
13		5	=C6/$H6	=D6/$H6	=E6/$H6	=F6/$H6	=G6/$H6	
14								
15			43	28	19	14	2	=SUM(C15:G15)
16	P_0		=C15/$H$15	=D15/$H$15	=E15/$H$15	=F15/$H$15	=G15/$H$15	
17								
18	P_1		=MMULT(C16:G16,C9:G13)	=MMULT(C16:G16,C9:G13)	=MMULT(C16:G16	=MMULT(C16:G16,C	=MMULT(C16:G16	=SUM(C18:G18)
19			=$H$15*C18	=$H$15*D18	=$H$15*E18	=$H$15*F18	=$H$15*G18	=SUM(C19:G19)

图 12-23　Excel 计算的部分公式



注意：计算 1 年后各类病人分布概率 $\boldsymbol{P}_1$ 是在选定单元格区域 C18:G18 情况下，在编辑栏输入公式“=MMULT(C16:G16,C9:G13)”，然后按 Ctrl+Shift+Enter 组合键产生的，其中单元格区域 C16:G16 为病人分类的初始状态概率行向量 $\boldsymbol{P}_0$，单元格区域 C9:G13 为转移概率矩阵 $\boldsymbol{P}$。

	A	B	C	D	E	F	G	H
1		类别	1	2	3	4	5	合计
2		1	146	51	3	0	0	200
3		2	47	75	28	1	0	151
4		3	2	17	65	17	0	101
5		4	0	3	9	31	37	80
6		5	0	0	0	0	60	60
7								592
8		类别	1	2	3	4	5	
9	P	1	0.730	0.255	0.015	0.000	0.000	
10		2	0.311	0.497	0.185	0.007	0.000	
11		3	0.020	0.168	0.644	0.168	0.000	
12		4	0.000	0.038	0.113	0.388	0.463	
13		5	0.000	0.000	0.000	0.000	1.000	
14								
15			43	28	19	14	2	106
16	P_0		0.406	0.264	0.179	0.132	0.019	
17								
18	P_1		0.3819	0.2698	0.1853	0.0831	0.0800	1
19			40	29	20	9	8	106

图 12-24　与图 12-23 相对应的计算结果

（2）2 年转移后的状态概率

$$\boldsymbol{P}_2=\boldsymbol{P}_0\boldsymbol{P}^2=(0.406,0.264,0.179,0.132,0.019)\begin{vmatrix}0.730 & 0.255 & 0.015 & 0 & 0\\ 0.311 & 0.497 & 0.185 & 0.007 & 0\\ 0.020 & 0.168 & 0.644 & 0.168 & 0\\ 0 & 0.038 & 0.113 & 0.388 & 0.463\\ 0 & 0 & 0 & 0 & 1\end{vmatrix}^2$$

$$=(0.3664,0.2657,0.1843,0.0652,0.1184)$$

采用 106 乘以各概率可预测得 2 年转移后的 1、2、3、4、5 类慢性肾病人数分别为 39、28、20、7、13。Excel 计算的关键是计算 $\boldsymbol{P}^2$，然后可仿照 1 年转移后的状态概率的计算获得结果。

单元格区域 C9:G13 为转移概率矩阵 $\boldsymbol{P}$，如果将 $\boldsymbol{P}^2$ 的计算结果放在单元格区域 C22:G26，选取 C22:G26，在编辑栏键入公式“=MMULT(C9:G13,C9:G13)”，然后按 Ctrl+Shift+Enter 组合键，得到：

$$\boldsymbol{P}^2=\begin{vmatrix}0.730 & 0.255 & 0.015 & 0 & 0\\ 0.311 & 0.497 & 0.185 & 0.007 & 0\\ 0.020 & 0.168 & 0.644 & 0.168 & 0\\ 0 & 0.038 & 0.113 & 0.388 & 0.463\\ 0 & 0 & 0 & 0 & 1\end{vmatrix}^2=\begin{vmatrix}0.613 & 0.315 & 0.068 & 0.004 & 0\\ 0.385 & 0.358 & 0.217 & 0.037 & 0.003\\ 0.080 & 0.203 & 0.465 & 0.175 & 0.078\\ 0.014 & 0.052 & 0.123 & 0.169 & 0.642\\ 0 & 0 & 0 & 0 & 1\end{vmatrix}$$

同理，单元格区域 C9:G13 为 $\boldsymbol{P}$，C22:G26 为 $\boldsymbol{P}^2$，如果将 $\boldsymbol{P}^3=\boldsymbol{P}\boldsymbol{P}^2$ 的计算结果放在单元格区域 L22:P26，选取 L22:P26，在编辑栏键入公式“=MMULT(C9:G13,C22:G26)”，然后按 Ctrl+Shift+Enter 组合键即可得到 $\boldsymbol{P}^3$，如图 12-25 所示（见光盘文件“12-10 Markov 预测.xlsx”）。

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
21		类别	1	2	3	4	5				类别	1	2	3	4	5	
22	P^2	1	0.613	0.315	0.068	0.004	0.000			P^3	1	0.547	0.324	0.112	0.015	0.002	
23		2	0.385	0.358	0.217	0.037	0.003				2	0.397	0.314	0.216	0.053	0.020	
24		3	0.080	0.203	0.465	0.175	0.078				3	0.131	0.206	0.358	0.147	0.159	
25		4	0.014	0.052	0.123	0.169	0.642				4	0.029	0.056	0.108	0.087	0.720	
26		5	0.000	0.000	0.000	0.000	1.000				5	0.000	0.000	0.000	0.000	1.000	
27																	
28	P_2		0.3664	0.2657	0.1843	0.0652	0.1184	1		P_3		0.3538	0.2589	0.1807	0.0580	0.1485	1
29			39	28	20	7	13	106				38	27	19	6	16	106

图 12-25　第 2、3 年各类慢性肾病患者的预测例数

（3）3 年转移后的状态概率

$$\boldsymbol{P}_3=\boldsymbol{P}_0\boldsymbol{P}^3=(0.406,0.264,0.179,0.132,0.019)\begin{vmatrix}0.730 & 0.255 & 0.015 & 0 & 0\\ 0.311 & 0.497 & 0.185 & 0.007 & 0\\ 0.020 & 0.168 & 0.644 & 0.168 & 0\\ 0 & 0.038 & 0.113 & 0.388 & 0.463\\ 0 & 0 & 0 & 0 & 1\end{vmatrix}^3$$

=(0.3538,0.2589,0.1807,0.0580,0.1485)

采用 106 乘以各概率可预测得 3 年后转移的 1、2、3、4、5 类慢性肾病病人数分别为 38、27、19、6、16。

12.6 灰色预测模型

黑色、白色之间的颜色为灰色，黑色表示情况完全未知，白色表示情况完全明了，灰色则是指情况处于是明非明了的状态。灰色预测模型也称灰色数列模型（grey dynamics model，GM），是对时间序列数据建立微分方程进行预测的一种统计学方法。GM(1，1)模型是灰色预测模型中最常用，也是灰色预测模型中最简单的一种预测模型，根据最小二乘法原理可求解一阶微分方程的参数。本节将结合实例介绍如何运用 Excel 进行 GM(1,1)灰色模型的预测与评价。

12.6.1 GM(1,1)模型的预测与评价方法

设原始数列排成时间数列 X_t（t=0,1,2,⋯,n），其中 X_0 为基准年份对应的原始数值。实现 GM(1,1)模型进行预测与评价的方法如下：

① 获得累加生成数列

通常原始数列数据没有明显的规律性，呈现零乱现象，为了使无规律的原始数据变得规律更为明显，可对原始数列进行累加生成处理，令累加生成的新变量为 Y_t（t=0,1,2,⋯,n），则有：

$$Y_t = \sum_{i=0}^{t} X_i \qquad t = 0,1,2,\cdots,n \tag{12-4}$$

② 获得累加生成数列的移动平均值

令移动平均值为 Z_t，则有：

$$Z_t = \frac{\left(Y_{t-1} + Y_t\right)}{2} \qquad t = 0,1,2,\cdots,n \tag{12-5}$$

③ 建立模型 GM(1,1)

Y_t 的一阶线性微分方程为：

$$\frac{\mathrm{d}Y_t}{\mathrm{d}t} + aY_t = \mu \qquad t = 0,1,2,\cdots,n \tag{12-6}$$

其中 α、μ 为待定参数，α 也称发展灰数，μ 也称内生控制灰数。按微分方程的求解方法，公式（12-6）也可写成：

$$Y_t=\left(X_0-\frac{\mu}{\alpha}\right)\mathrm{e}^{-\alpha t}+\frac{\mu}{\alpha}\qquad t=0,1,2,\cdots,n \tag{12-7}$$

其中 X_0 为基准年份对应的原始数值。根据最小二乘法原理可以得到待定参数：

$$\alpha=\frac{\sum_{t=1}^{n}X_t\sum_{t=1}^{n}Z_t-n\sum_{t=1}^{n}X_tZ_t}{n\sum_{t=1}^{n}Z_t^2-\left(\sum_{t=1}^{n}Z_t\right)^2} \tag{12-8}$$

$$\mu=\frac{\sum_{t=1}^{n}X_t\sum_{t=1}^{n}Z_t^2-\sum_{t=1}^{n}X_tZ_t\sum_{t=1}^{n}Z_t}{n\sum_{t=1}^{n}Z_t^2-\left(\sum_{t=1}^{n}Z_t\right)^2} \tag{12-9}$$

④求预测值

可通过公式（12-8）、（12-9）计算获得两个参数 α 、μ 的估计值，然后将 α 和 μ 的估计值代入公式（12-7）获得累加生成数列的估计值 $\hat{Y}_t$，原始数列值 X_t 的预测估计值为：

$$\hat{X}_t=\hat{Y}_t-\hat{Y}_{t-1}\qquad t=1,2,\cdots,n \tag{12-10}$$

⑤灰色模型拟合情况的评价

①计算绝对误差 δ_t 及其总体标准差 S_2，可用公式：

$$\delta_t=X_t-\hat{X}_t\qquad t=1,2,\cdots,n$$

$$S_2=\sqrt{\frac{\sum_{t=1}^{n}\left(\delta_t-\bar{\delta}\right)^2}{n}} \tag{12-11}$$

②计算原始数列 X_t（t=1,2,⋯,n）的总体标准差

$$S_1=\sqrt{\frac{\sum_{t=1}^{n}\left(X_t-\bar{X}\right)^2}{n}} \tag{12-12}$$

③计算后验差比值 C

$$C=\frac{S_2}{S_1} \tag{12-13}$$

④计算小误差概率 P

$$P=P\left\{\left|\delta_t-\bar{\delta}\right|<0.6745S_1\right\}=\sum k/n \tag{12-14}$$

如果$\left|\delta_t-\bar{\delta}\right|<0.6745S_1$则 k=1，否则 k=0。清点 k=1 的数量（$\sum k$）所占年份数（n）的比例，即获得小误差概率 P。

⑤结果判断

拟合优度分成 4 个等级，具体判断如表 12-5 所示。

表 12-5　灰色模型拟合优度的判断

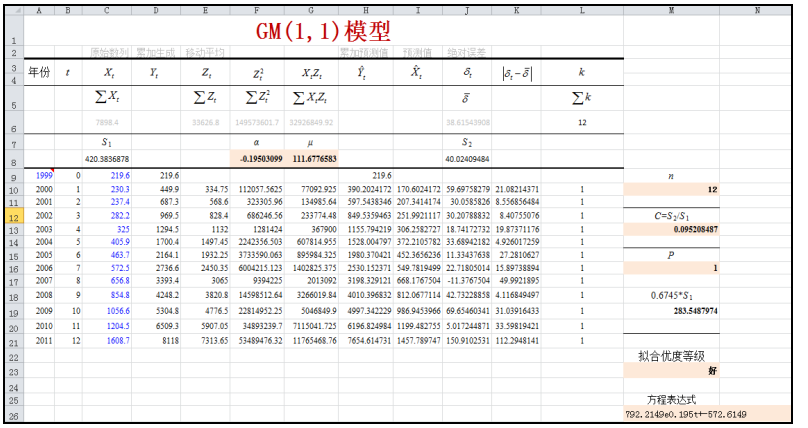
拟合优度等级	P 值	C 值
好	>0.95	<0.35
合格	>0.80	<0.50
勉强	>0.70	<0.65
不合格	≤0.70	≥0.65

⑥事前与事中模型适应性检验

以 X_{t-1}/X_t（t=1,2,⋯,n）的值大于等于 0.1353，且小于等于 7.389 为“好”，如果“好”的百分率在 90%以上，表示资料适合进行灰色模型预测。也可比较相对误差 $\dfrac{X_t-\hat{X}_t}{X_t}$（$t$=1,2,⋯,$n$）是否均较小，值越小越适合进行灰色模型预测。

12.6.2　灰色模型预测实例

例 12-6：图 12-26 的 A9:C21 单元格区域（见光盘文件“12-11 灰色预测模型.xlsx”）列出了湖北省 1999~2011 年的人均卫生费用数据。试建立灰色预测模型，并预测该省后续两年的年人均卫生总费用。



GM(1,1)模型

年份	t	X_t	Y_t	Z_t	Z_t^2	X_tZ_t	$\hat{Y}_t$	$\hat{X}_t$	δ_t	$\lvert\delta_t-\bar{\delta}\rvert$	k
		$\sum X_t$		$\sum Z_t$	$\sum Z_t^2$	$\sum X_tZ_t$			$\bar{\delta}$		$\sum k$
		7898.4		33626.8	149573601.7	32926849.92			38.61543908		12
		S_1			α	μ			S_2		
		420.3836878			-0.19503099	111.6776583			40.02409484		
1999	0	219.6	219.6				219.6				
2000	1	230.3	449.9	334.75	112057.5625	77092.925	390.2024172	170.6024172	59.69758279	21.08214371	1
2001	2	257.4	687.3	568.6	323305.96	134985.64	597.5438346	207.3414174	30.0585826	8.556856484	1
2002	3	282.2	969.5	828.4	686246.56	233774.48	849.5359463	251.9921117	30.20788832	8.40755076	1
2003	4	325	1294.5	1132	1281424	367900	1155.794219	306.2582727	18.74172732	19.87371176	1
2004	5	405.9	1700.4	1497.45	2242356.503	607814.955	1528.004797	372.2105782	33.68942182	4.926017259	1
2005	6	463.7	2164.1	1932.25	3733590.063	895984.325	1980.370421	452.3656236	11.33437638	27.2810627	1
2006	7	572.5	2736.6	2450.35	6004215.123	1402825.375	2530.152371	549.7819499	22.71805014	15.89738894	1
2007	8	656.8	3393.4	3065	9394225	2013092	3198.329121	668.1767504	-11.3767504	49.9921895	1
2008	9	854.8	4248.2	3820.8	14598512.64	3266019.84	4010.396832	812.0677114	42.73228858	4.116849497	1
2009	10	1056.6	5304.8	4776.5	22814952.25	5046849.9	4997.342229	986.9453966	69.65460341	31.03916433	1
2010	11	1204.5	6509.3	5907.05	34893239.7	7115041.725	6196.824984	1199.482755	5.017244871	33.59819421	1
2011	12	1608.7	8118	7313.65	53489476.32	11765468.76	7654.614731	1457.789747	150.9102531	112.2948141	1

n	12
$C=S_2/S_1$	0.095208487
P	1
$0.6745*S_1$	283.5487974
拟合优度等级	好
方程表达式	792.2149e0.195t−572.6149

图 12-26　GM（1,1）模型的数据及主要计算结果

解：按 12.6.1 节所给方法编制 GM（1,1）模型的计算表，在图 12-26 界面，选择功能区“公式→公式审核→显示公式”命令，可查看图 12-26 中相应的 Excel 公式，部分 Excel 公式如图 12-27 所示。

	A	B	C	D	E	F	G	H	I	J
1					GM(1,1)模型					
2			原始数列	累加生成	移动平均			累加预测值	预测值	绝对误差
3	年份	t	X_t	Y_t	Z_t	Z_t^2	X_tZ_t	$\hat{Y}_t$	$\hat{X}_t$	ε_t
4										
5			$\sum X_t$		$\sum Z_t$	$\sum Z_t^2$	$\sum X_tZ_t$			$\bar{\varepsilon}$
6			=SUM(C10:C60)		=SUM(E10:E60)	=SUM(F10:F60)	=SUM(G10:G60)			=AVERAGE(J10:J60)
7			S_1			a	μ			S_2
8			=STDEV.P(C10:C60)			=(C6*F6-M10*G6)/(M10*F6-I =(C6*F6-G6*E6)/(M10*F6-E6				=STDEV.P(J10:J60)
9	1999	0	219.6	=C9				=IF(C9<>"",($C$9-$G$8/$F$8)		
10	=IF(C10<>"",A9	=IF(C10<>"",B	230.3	=IF(C10<>"",D9+C10,"")	=IF(C10<>"",(D9−D10)/2,"	=IF(C10<>"",E10^2,"")	=IF(C10<>"",C10*E10,"")	=IF(C10<>"",($C$9-$G$8/$F$8=IF(C10<>"",H10-H9,"")		=IF(C10<>"",C10-I10
11	=IF(C11<>"",A1	=IF(C11<>"",B	237.4	=IF(C11<>"",D10+C11,"")	=IF(C11<>"",(D10+D11)/2,	=IF(C11<>"",E11^2,"")	=IF(C11<>"",C11*E11,"")	=IF(C11<>"",($C$9-$G$8/$F$8=IF(C11<>"",H11-H10,"")		=IF(C11<>"",C11-I11
12	=IF(C12<>"",A1	=IF(C12<>"",B	282.2	=IF(C12<>"",D11+C12,"")	=IF(C12<>"",(D11+D12)/2,	=IF(C12<>"",E12^2,"")	=IF(C12<>"",C12*E12,"")	=IF(C12<>"",($C$9-$G$8/$F$8=IF(C12<>"",H12-H11,"")		=IF(C12<>"",C12-I12
13	=IF(C13<>"",A1	=IF(C13<>"",B	325	=IF(C13<>"",D12+C13,"")	=IF(C13<>"",(D12+D13)/2,	=IF(C13<>"",E13^2,"")	=IF(C13<>"",C13*E13,"")	=IF(C13<>"",($C$9-$G$8/$F$8=IF(C13<>"",H13-H12,"")		=IF(C13<>"",C13-I13
14	=IF(C14<>"",A1	=IF(C14<>"",B	405.9	=IF(C14<>"",D13+C14,"")	=IF(C14<>"",(D13+D14)/2,	=IF(C14<>"",E14^2,"")	=IF(C14<>"",C14*E14,"")	=IF(C14<>"",($C$9-$G$8/$F$8=IF(C14<>"",H14-H13,"")		=IF(C14<>"",C14-I14
15	=IF(C15<>"",A1	=IF(C15<>"",B	463.7	=IF(C15<>"",D14+C15,"")	=IF(C15<>"",(D14+D15)/2,	=IF(C15<>"",E15^2,"")	=IF(C15<>"",C15*E15,"")	=IF(C15<>"",($C$9-$G$8/$F$8=IF(C15<>"",H15-H14,"")		=IF(C15<>"",C15-I15
16	=IF(C16<>"",A1	=IF(C16<>"",B	572.5	=IF(C16<>"",D15+C16,"")	=IF(C16<>"",(D15+D16)/2,	=IF(C16<>"",E16^2,"")	=IF(C16<>"",C16*E16,"")	=IF(C16<>"",($C$9-$G$8/$F$8=IF(C16<>"",H16-H15,"")		=IF(C16<>"",C16-I16
17	=IF(C17<>"",A1	=IF(C17<>"",B	656.8	=IF(C17<>"",D16+C17,"")	=IF(C17<>"",(D16+D17)/2,	=IF(C17<>"",E17^2,"")	=IF(C17<>"",C17*E17,"")	=IF(C17<>"",($C$9-$G$8/$F$8=IF(C17<>"",H17-H16,"")		=IF(C17<>"",C17-I17
18	=IF(C18<>"",A1	=IF(C18<>"",B	854.8	=IF(C18<>"",D17+C18,"")	=IF(C18<>"",(D17+D18)/2,	=IF(C18<>"",E18^2,"")	=IF(C18<>"",C18*E18,"")	=IF(C18<>"",($C$9-$G$8/$F$8=IF(C18<>"",H18-H17,"")		=IF(C18<>"",C18-I18
19	=IF(C19<>"",A1	=IF(C19<>"",B	1056.6	=IF(C19<>"",D18+C19,"")	=IF(C19<>"",(D18+D19)/2,	=IF(C19<>"",E19^2,"")	=IF(C19<>"",C19*E19,"")	=IF(C19<>"",($C$9-$G$8/$F$8=IF(C19<>"",H19-H18,"")		=IF(C19<>"",C19-I19
20	=IF(C20<>"",A1	=IF(C20<>"",B	1204.5	=IF(C20<>"",D19+C20,"")	=IF(C20<>"",(D19+D20)/2,	=IF(C20<>"",E20^2,"")	=IF(C20<>"",C20*E20,"")	=IF(C20<>"",($C$9-$G$8/$F$8=IF(C20<>"",H20-H19,"")		=IF(C20<>"",C20-I20
21	=IF(C21<>"",A2	=IF(C21<>"",B	1608.7	=IF(C21<>"",D20+C21,"")	=IF(C21<>"",(D20+D21)/2,	=IF(C21<>"",E21^2,"")	=IF(C21<>"",C21*E21,"")	=IF(C21<>"",($C$9-$G$8/$F$8=IF(C21<>"",H21-H20,"")		=IF(C21<>"",C21-I21
22	=IF(C22<>"",A2	=IF(C22<>"",B		=IF(C22<>"",D21+C22,"")	=IF(C22<>"",(D21+D22)/2,	=IF(C22<>"",E22^2,"")	=IF(C22<>"",C22*E22,"")	=IF(C22<>"",($C$9-$G$8/$F$8=IF(C22<>"",H22-H21,"")		=IF(C22<>"",C22-I22
23	=IF(C23<>"",A2	=IF(C23<>"",B		=IF(C23<>"",D22+C23,"")	=IF(C23<>"",(D22+D23)/2,	=IF(C23<>"",E23^2,"")	=IF(C23<>"",C23*E23,"")	=IF(C23<>"",($C$9-$G$8/$F$8=IF(C23<>"",H23-H22,"")		=IF(C23<>"",C23-I23
24	=IF(C24<>"",A2	=IF(C24<>"",B		=IF(C24<>"",D23+C24,"")	=IF(C24<>"",(D23+D24)/2,	=IF(C24<>"",E24^2,"")	=IF(C24<>"",C24*E24,"")	=IF(C24<>"",($C$9-$G$8/$F$8=IF(C24<>"",H24-H23,"")		=IF(C24<>"",C24-I24
25	=IF(C25<>"",A2	=IF(C25<>"",B		=IF(C25<>"",D24+C25,"")	=IF(C25<>"",(D24+D25)/2,	=IF(C25<>"",E25^2,"")	=IF(C25<>"",C25*E25,"")	=IF(C25<>"",($C$9-$G$8/$F$8=IF(C25<>"",H25-H24,"")		=IF(C25<>"",C25-I25
26	=IF(C26<>"",A2	=IF(C26<>"",B		=IF(C26<>"",D25+C26,"")	=IF(C26<>"",(D25+D26)/2,	=IF(C26<>"",E26^2,"")	=IF(C26<>"",C26*E26,"")	=IF(C26<>"",($C$9-$G$8/$F$8=IF(C26<>"",H26-H25,"")		=IF(C26<>"",C26-I26

图 12-27　GM（1,1）模型的部分 Excel 公式

再重复单击“公式→公式审核→显示公式”命令，可恢复到图 12-26 的正常数据结果界面。

该程序获得类似公式（12-7）的 GM（1,1）灰色预测模型为：

$$\hat{Y}_t = 792.2149\mathrm{e}^{0.1950t} - 572.6149 \qquad t = 0,1,2,\ldots,n \qquad (12\text{-}15)$$

只要将 t 代入该公式，便可获得累加生成值的预测值。例如，预测 2012、2013 两年的湖北省年人均卫生费用，将 t=13、14（如图 12-26 所示）值代入公式（12-15），得到累加生成预测值为 9422、11574；由公式（12-10）获得对应预测值为 1770、1574。实际上光盘程序 12-12 已经自动计算了 5 年的预测值（更精确），如图 12-28 所示。注意，因为序列较短，预测两年尚可，如预测的年份越靠未来，预测值会越不准确。所以程序尽管给出 5 年，读者可向后预测两年或三年即可。

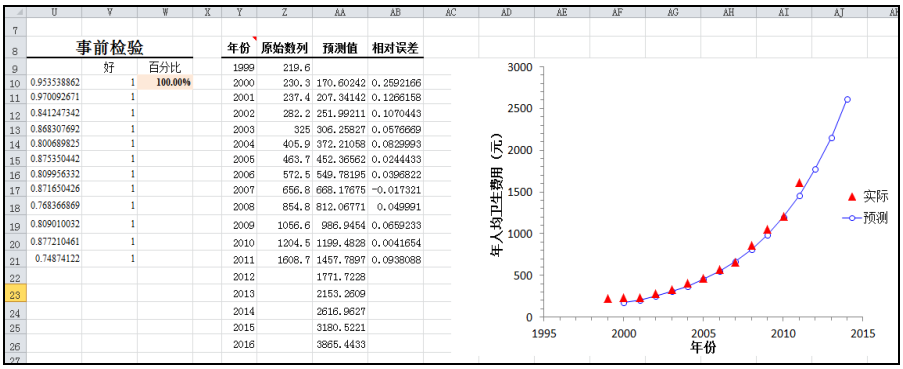
	M	N	O	P	Q	R	S
25	方程表达式						
26	792.2149e0.195t+-572.6149			注：方程中出现“+-”表示减“-”！！			
27							
28	年份	$\hat{X}_t$ 预测值	$\hat{Y}_t$ 预测值				
29	2012	1771.7228	9426.337532				
30	2013	2153.260913	11579.59844				
31	2014	2616.962744	14196.56119				
32	2015	3180.522137	17377.08333				
33	2016	3865.443285	21242.52661				

图 12-28　GM（1,1）模型的预测值

本例的后验差比值 C= 0.0952<0.35，小误差概率 P= 1>0.95（如图 12-26 所示），表明该

模型预测为“好”（参见对照表 12-5）。

事前检验“好”，百分率为 100%，事中检验得到“相对误差”，绝大多数（除 2000 年与 2001 年）在 10%及以内（如图 12-29 所示），同样表明模型拟合尚可。



事前检验		
	好	百分比
0.953538862	1	100.00%
0.970092671	1	
0.841247342	1	
0.868307692	1	
0.800689825	1	
0.875350442	1	
0.809956332	1	
0.871650426	1	
0.768366869	1	
0.809010032	1	
0.877210461	1	
0.74874122	1	

年份	原始数列	预测值	相对误差
1999	219.6		
2000	230.3	170.60242	0.2592166
2001	237.4	207.34142	0.1266158
2002	282.2	251.99211	0.1070443
2003	325	306.25827	0.0576669
2004	405.9	372.21058	0.0829993
2005	463.7	452.36662	0.0244433
2006	572.5	549.78195	0.0396822
2007	656.8	668.17675	-0.017321
2008	854.8	812.06771	0.049991
2009	1056.6	986.9454	0.0659233
2010	1204.5	1199.4828	0.0041654
2011	1608.7	1457.7897	0.0938088
2012		1771.7228	
2013		2153.2609	
2014		2616.9627	
2015		3180.5221	
2016		3865.4433	

图 12-29　事前与事中的模型适应性判断

温馨提示：用户只需修改图 12-26 中的单元格 A9（基准年份）以及单元格区域 C9:C60 范围内的原始数列数据，工作表便可自动完成 GM（1,1）灰色预测模型的绝大多数计算。为了正确得到图 12-29 中右侧的预测趋势图，需要根据实际数据修改单元格区域 Y9:AB60 的链接，以及趋势图的“选择数据”选项。

注意：该程序设计最多可容纳 50 个原始数列数据。如果有更多原始数列数据，读者可采用密码“0”，撤销工作表保护后对程序进行适当修改。

12.7　傅利叶分析

时间序列分析包括两个范畴：时间域分析和频率域分析。时域函数经傅利叶变换后自变量由时间变为频率，使得对原始序列中包含的周期活动成分的研究成为可能；同时，在频率域经过处理的频域函数经傅利叶逆变换又可转换回时间域，离散时间序列是连续时域信号的采样结果，也是实际工作中观察得到的结果。

离散时域函数的傅利叶变换及频域函数的傅利叶逆变换可以分别表示为：

$$X_k = \frac{1}{N}\sum_{t=0}^{N-1} x_t \mathrm{e}^{-j2\pi tk/N} \quad (k=0,1,2,\cdots)$$

$$x_t = \sum_{k=0}^{N-1} X_k \mathrm{e}^{j2\pi tk/N} \quad (k=0,1,2,\cdots)$$

从理论上讲，N 可以是任何自然数，但基于快速傅利叶变换的运算实现要求 $N=2^m(m=1,2,\cdots)$，式中 $j=1,2,\cdots$，π 为圆周率，Excel 中对数据也有这样的限定。

例 12-7：设离散时域函数的一系列观察值为例 12-1 中 1994～2001 年的发病率，试求该时域序列的傅利叶变换结果，然后再将所得结果变换回时间域。

解：采用 Excel 数据分析工具——傅利叶分析的具体操作步骤如下：

①将图 12-1 的单元格区域 A9:B16 的数据复制、粘贴到新的工作表中，如图 12-30 所示的单元格区域 A3:B11（见光盘文件“12-12 傅利叶分析.xlsx”）。

②选择“数据→分析→数据分析→傅利叶分析”命令，单击“确定”按钮。

③在弹出的“傅利叶分析”对话框（如图 12-30 右侧所示）中，输入区域选择 B3:B11，选择“标志位于第一行”复选框，输出区域选择单元格 D4，单击“确定”按钮。

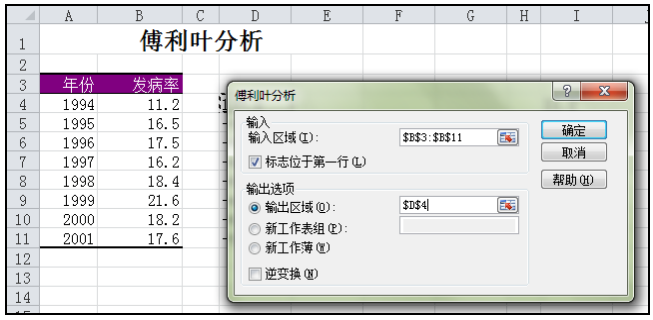


图 12-30　傅利叶变换的原始数据与操作选项

④在单元格区域 D4:D11 中所得结果即为傅利叶变换结果，如图 12-31 所示（见光盘文件“12-12 傅利叶分析.xlsx”）。注意：结果中的 i 是虚数符号，即 i 的平方为“-1”，傅利叶变换涉及复数的有关计算。



图 12-31　傅利叶分析结果

进一步对单元格区域 D4:D11 的频率域函数值作傅利叶逆变换，具体操作步骤如下：

①选中功能区“数据→分析→数据分析→傅利叶分析”后，单击“确定”命令。

② 在弹出的“傅利叶分析”对话框中，输入区域选择 D4:D11，不选择“标志位于第一行”复选框；输出区域选择单元格 I4；选择“逆变换”复选框；单击“确定”按钮。

③ 在单元格区域 I4:I11 中获得的数据即为傅利叶逆变换结果，其结果与原始序列 B4:B11 相同，如图 12-31 所示。对傅利叶变换结果作逆变换，获得的结果为原序列本身。

12.8 圆形分布分析

在第 12.4 节介绍了季节预测法。类似于季节，很多资料具有周而复始的特征，如每年的月份、每天的时间、一个圆的角度等。他们没有真正的零点，也没有数值大小的区分。关于圆角度的实例有：肛门上痔疮的位置，子宫颈口溃疡的位置，眼角膜上溃疡疤痕的位置，乳房上肿块的位置等。凡呈圆形的器官或者组织上发现有异常之处，都可当作一个圆周标出其位置。对于这类资料，不能按常规方法处理，而应该进行圆形分布分析。

12.8.1 角度变换

圆形分布方法的基本思想是：将所有周而复始、无数值大小的各种数值转变成角度值。再将角度转变成弧度值后，利用正弦、余弦等函数进一步求出平均角、角标准差等。用集中测度 r 值进行假设检验。最后再把角度换算回原始尺度。

以时钟 0 点（正北）为圆形的 0°，以时钟 3 点（正东）为圆形的 90°，以 6 点（正南）为圆形的 180°，以 9 点（正西）为圆形的 270°，12 点与 0 点相重叠处也是圆形的 360°（见图 12-32 左图）；昼夜时间（以 24 小时计）点 0:00、6:00、12:00、1800、24:00 点（或 0:00 点）（见图 12-32 右图），分别对应圆形的 0°、90°、180°、270°、360°；每年的 12 个月分别对应圆形的 30°、60°、90°、120°、150°、180°、210°、240°、270°、300°、330°、360°。其他依次类推。自 0° 开始，顺时针方向为正，逆时针方向为负。具体角度变换参见光盘文件“12-13 圆形分布分析.xlsx”的“角度变换”工作表。

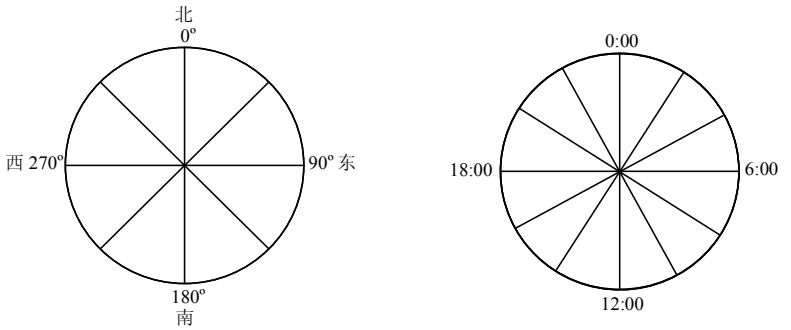


图 12-32　圆形位置对应的角度

12.8.2 平均角及其假设检验

设 a_i 表示 $\alpha_1,\alpha_2\cdots,\alpha_n$（$i=1,2,\cdots,n$）的 n 个角，$\overline{a}$ 表示平均角；f_i 表示第 i 角的频数；$\overline{x}$、$\overline{y}$ 分别为平均向量的横坐标、纵坐标。r 为平均向量矢量，表示集中测度，无单位，取值在 0～1。当 $r=0$ 时表示没有一个平均方向，即角度呈随机分布；当 $r=1$ 时表示全部数据都集中在同一个方向。上述指标相应的计算公式为：

$$x_i=\cos\left(\alpha_i\frac{\pi}{180}\right)\quad y_i=\sin\left(\alpha_i\frac{\pi}{180}\right)\quad n=\sum f_i$$

$$\overline{x}=\left(\sum f_i x_i\right)\Big/ n\quad \overline{y}=\left(\sum f_i y_i\right)\Big/ n\quad r=\sqrt{\overline{x}^2+\overline{y}^2}$$

与集中测度 r 对应的离散测度 s，即角标准差，其计算公式为：

$$s=\frac{180}{\pi}\sqrt{-2\ln r}$$

集中测度 r 越大，离散测度 s 越小，表示角的数值越集中，反之角越分散。离散测度 s 的取值范围为 0～∞。

平均角 $\overline{a}$ 对于不同的 $\overline{x}$、$\overline{y}$ 有不同的取值，即：

$$\overline{a}=\begin{cases}\dfrac{180}{\pi}\operatorname{arctg}(\overline{y}/\overline{x}) & \text{当}\overline{x}>0\\ 180+\dfrac{180}{\pi}\operatorname{arctg}(\overline{y}/\overline{x}) & \text{当}\overline{x}<0\\ 90 & \text{当}\overline{x}=0,\text{且}\overline{y}>0\\ 270 & \text{当}\overline{x}=0,\text{且}\overline{y}<0\\ \text{不定} & \text{当}\overline{x}=0,\text{且}\overline{y}=0\end{cases}$$

平均角对应的正弦、余弦、正切分别为 $\sin\overline{a}=\overline{y}/r$、$\cos\overline{a}=\overline{x}/r$、$\operatorname{tg}\overline{a}=\overline{y}/\overline{x}$。正弦、余弦的正负符号决定了平均角所在象限的位置，正弦、余弦均为正值，则角位于第一象限；正弦为正、余弦为负，则角位于第二象限；正弦、余弦均为负值，则角位于第三象限；正弦为负、余弦为正，则角位于第四象限（如图 12-33 所示）。

检验所有角度 a_i 均匀分布在圆周，即检验样本集中测度 r 对应的总体集中测度 $\rho=0$，可采用雷氏检验（Rayleigh's test）。先采用公式：

$$Z=nr^2$$

计算雷氏 Z 值，并与雷氏 Z 临界值（见光盘文件“12-13 圆形分布分析.xlsx”的“圆形分

布分析”工作表右侧）进行比较。在 $\alpha = 0.05$ 水准，雷氏 Z 临界值随例数 n 变化而变化，n 越大临界值越大，如 n=6，$Z_{0.05} = 2.865$；n=∞，$Z_{0.05} = 2.99$。此临界值变化并不大，一般在 2.9 左右。

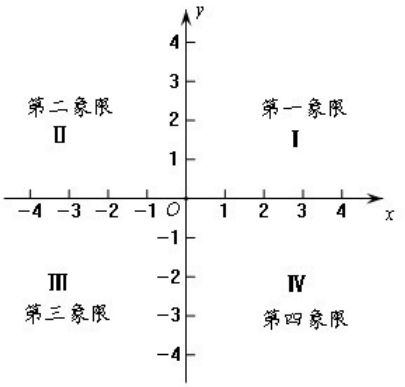


图 12-33　象限位置

采用公式

$$\overline{a} \pm \sin^{-1}\left(z_{0.05/2}\sqrt{\frac{1-a2}{2nr^2}} \right)，\text{其中 } a2 = \frac{\sum_{i=1}^{n} f_i \cos\left[\frac{2\pi}{180}\left(a_i - \overline{a}\right)\right]}{n}$$

可获得总体平均角的 95%置信区间（张福林，杨树勤. 中国卫生统计，1992，9(2)：20）。

12.8.3　圆形分布分析实例

按照上面介绍的方法已制作了光盘文件“12-13 圆形分布分析.xlsx”，文件中有三个主要工作表，一个为“角度变换”，另两个为“昼夜时点”与“月份”的圆形分布分析。

例 12-8：对某一马来丝虫感染者用等量血液涂片做微丝蚴计数，在昼夜时点 4、8、12、16、20、24（每 4 小时采血 1 次）观察到的微丝蚴条数分别为 91、13、7、70、151、161（摘自金丕焕，陈峰主编. 医用统计方法（第三版），2009:330），试用圆形分布分析微丝蚴的活动情况。

解：因为属于昼夜 24 小时资料，采用光盘文件 12-13 的“1. 一天 24 小时时间转换为圆周角度”进行角度变换，即在“角度变换”工作表的单元格区域 B5:B10 中输入“4、8、12、16、20、24”，将单元格区域 D5:D10 获得的角度值复制，选择性粘贴“数值”至工作表“昼夜时点”的单元格区域 B5:B10 中。在单元格区域 E5:E10 中依次键入“91、13、7、

70、151、161”，即可获得如图 12-34 所示界面的结果。

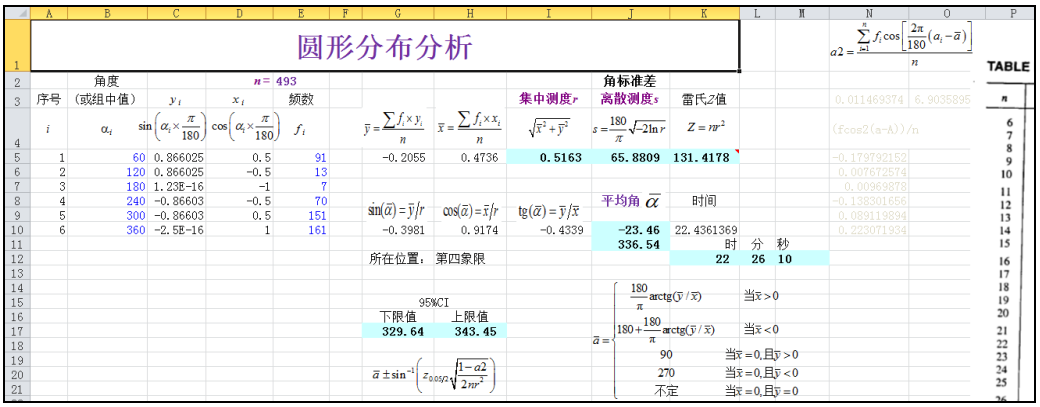


图 12-34 昼夜时点的圆形分布分析

图 12-34 的第 B 列和第 E 列是实际数据键入区，主要结果集中在第 G 至第 M 列，最右侧是雷氏 Z 临界值表。本例集中测度 r=0.5163，离散测度 s=65.8809，经雷氏检验，雷氏 Z=131.4178，远高于临界值 2.99，说明有平均角存在。本例平均角为 336.54°（95% CI 为 326.64°～343.45°），对应的昼夜时间是 22 时 26 分 10 秒。

例 12-9：2000～2001 年辽宁省大连金州区庄河地区卫生局提供了 66 例自杀患者的月份分布情况（胡冬梅，等.中国卫生统计，2004），如表 12-6 所示。试用圆形分布分析自杀的月份特征。

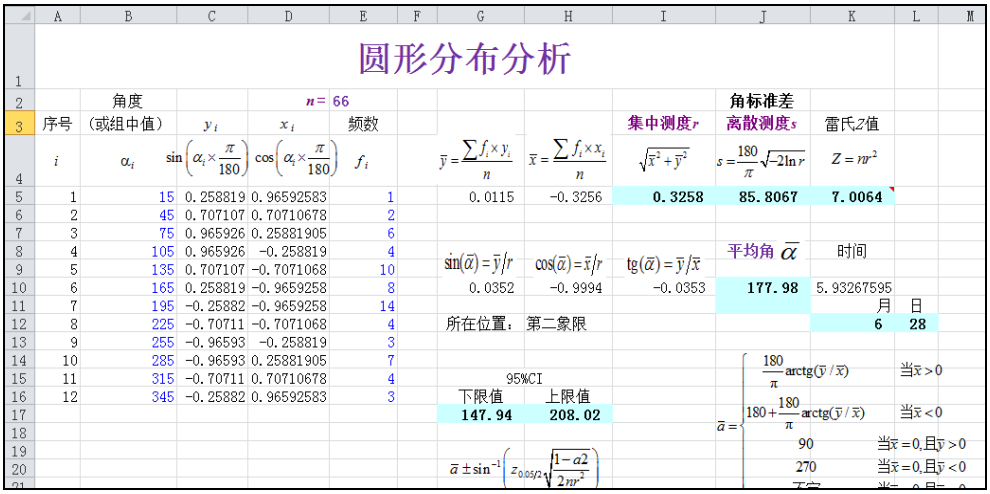
表 12-6 66 例自杀者自杀月份分布

月　份	1	2	3	4	5	6	7	8	9	10	11	12
人　数	1	2	6	4	10	8	14	4	3	7	4	3

解：因为属于月份资料，采用光盘文件中的“3. 1 年 12 个月为圆周”进行角度变换。将角度和频数资料在工作表“月份”相应位置输入后，可获得如图 12-35 所示界面的结果。

图 12-35 的第 B 列和第 E 列是实际数据键入区，主要结果集中在第 G 至第 L 列，最右侧是雷氏 Z 临界值表。本例集中测度 r=0.3258，离散测度 s=85.8067，经雷氏检验，雷氏 Z=7.0064，大于临界值 2.99，说明有平均角存在。本例平均角为 177.98°，对应的时间是 6 月 28 日，即自杀主要集中在 6 月 28 日左右。

温馨提示：对于光盘文件 12-13，每次只需输入角度和频数数据便可获得相关结果。如果为非频数数据，只需将频数输入数字填入“1”即可。该程序角度和频数的输入限制在 30 个以内。



圖形分布分析

序号	角度（或组中值）	y_i	x_i	频数
i	α_i	$\sin\left(\alpha_i\times\frac{\pi}{180}\right)$	$\cos\left(\alpha_i\times\frac{\pi}{180}\right)$	f_i
1	15	0.258819	0.96592583	1
2	45	0.707107	0.70710678	2
3	75	0.965926	0.25881905	6
4	105	0.965926	-0.258819	4
5	135	0.707107	-0.7071068	10
6	165	0.258819	-0.9659258	8
7	195	-0.25882	-0.9659258	14
8	225	-0.70711	-0.7071068	4
9	255	-0.96593	-0.258819	3
10	285	-0.96593	0.25881905	7
11	315	-0.70711	0.70710678	4
12	345	-0.25882	0.96592583	3

$n=66$

$\bar{y}=\frac{\sum f_i\times y_i}{n}$　$\bar{x}=\frac{\sum f_i\times x_i}{n}$　$\sqrt{\bar{x}^2+\bar{y}^2}$　$s=\frac{180}{\pi}\sqrt{-2\ln r}$　$Z=nr^2$

0.0115	-0.3256	0.3258	85.8067	7.0064

$\sin(\bar{\alpha})=\bar{y}/r$　$\cos(\bar{\alpha})=\bar{x}/r$　$\mathrm{tg}(\bar{\alpha})=\bar{y}/\bar{x}$

0.0352	-0.9994	-0.0353

所在位置：第二象限

95%CI

下限值	上限值
147.94	208.02

$\bar{\alpha}\pm\sin^{-1}\left(z_{0.05/2}\sqrt{\frac{1-a2}{2nr^2}}\right)$

角标准差离散测度 s　雷氏 Z 值

平均角 $\bar{\alpha}$　时间

177.98　5.93267595

月　日

6　28

$$\bar{\alpha}=\begin{cases}\frac{180}{\pi}\arctg(\bar{y}/\bar{x}) & \text{当}\bar{x}>0\\ 180+\frac{180}{\pi}\arctg(\bar{y}/\bar{x}) & \text{当}\bar{x}<0\\ 90 & \text{当}\bar{x}=0,\text{且}\bar{y}>0\\ 270 & \text{当}\bar{x}=0,\text{且}\bar{y}<0\end{cases}$$

图 12-35　月份的圆形分布分析

12.8.4　两个或多个样本平均角的比较

两个或多个样本的平均角各自经均匀性检验，如果都拒绝 H_0，则可用 Watson-William 检验来判断各平均角是否来自总体平均角为 ρ 的总体，即可比较平均角间的差别有无统计学意义。

设有 k 个样本，以 n_j、r_j 分别表示第 j 个样本（$j=1,2,\cdots,k$）的样本含量、集中测度，检验样本集中测度 r_j 对应的总体集中测度 ρ_j 是否相等的 Watson-William 检验是：

H_0：$\rho_1=\rho_2=\cdots=\rho_k$

H_1：ρ_j 不完全相等

$$F=K\frac{(N-k)\left(\sum_{j=1}^{k}R_j-R\right)}{(k-1)\left(N-\sum_{j=1}^{k}R_j\right)}\qquad \nu_1=k-1,\quad \nu_2=N-k$$

公式中 K 为与集中测度 r 有关的校正因子，r 越小 K 越大，如 $r=0.01$ 的 $K=19.75$；$r=1$ 为集中测度最大值，此时 K=1（见光盘文件“12-14 多样本比较的圆形分布分析.xlsx”中的“校正因子 K”工作表）。其中 $N=\sum n_j$，$R_j=n_jr_j$，$R=Nr$。

这里提到的 r 是合并集中测度，即将 k 个样本当成一个整体获取的集中测度值，计算公式为：

$$r=\sqrt{\bar{y}^2+\bar{x}^2}=\sqrt{\left(\frac{\sum_{j=1}^{k}\sum_{i=1}^{n}f_{ij}\sin\frac{\pi}{180}\alpha_{ij}}{N}\right)^2+\left(\frac{\sum_{j=1}^{k}\sum_{i=1}^{n}f_{ij}\cos\frac{\pi}{180}\alpha_{ij}}{N}\right)^2}$$

可利用 Excel 函数 FDIST 或 F.DIST 获得 F 值对应的 P 值。$P\leqslant 0.05$ 表示各平均角中，至少两组间差异有统计学意义。

注意：该方法要求各平均角必须经均匀性检验认为有意义才能进行比较，并且合并集中测度 r 须大于 0.45 时效果才较满意。

例 12-10：A 病室有 6 位病人，晚上入睡时间分别为 20:30、21:00、21:15、21:20、21:45、22:00；B 病室有 7 人，晚上入睡时间分别为 21:30、21:45、22:05、22:15、22:20、22:45、22:50（摘自金丕焕，陈峰主编. 医用统计方法（第三版），2009:331），问两病室病人入睡时间的早晚有无差别？

解：因为属于昼夜 24 小时资料，采用光盘文件中的“1. 一天 24 小时时间转换为圆周角度”进行角度转换。对于 A 室变换得到的角度为“307.50、315.00、318.75、320.00、326.25、330.00”，以同样的方法得到 B 室的角度变换值为“322.50、326.25、331.25、333.75、335.00、341.25、342.50”（见光盘文件“12-14 多样本比较的圆形分布分析.xlsx”中的“角度变换”工作表）。

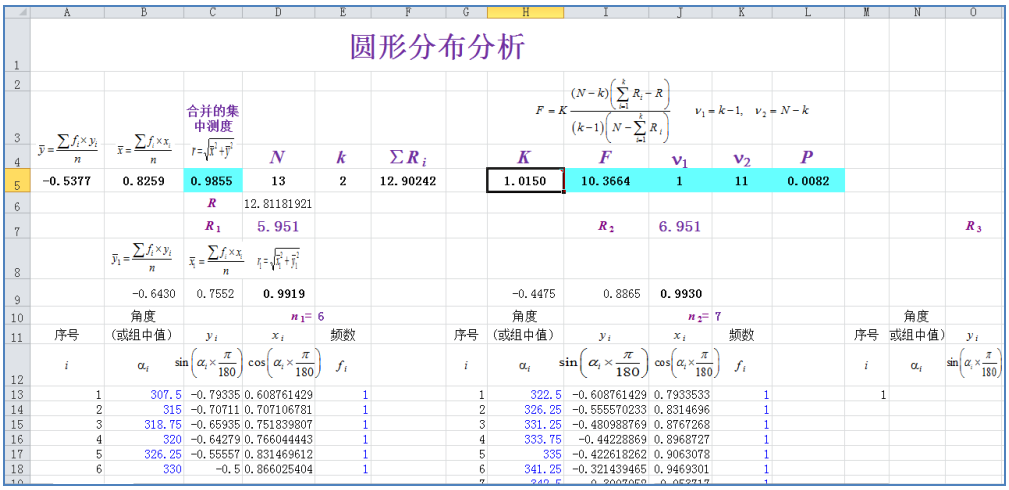
光盘文件“12-14 多样本比较的圆形分布分析.xlsx”是按上面提到的多样本圆形分布分析方法而制作，该程序适合 $k=2$ 至 5 的多样本比较。只需要将各样本的角度 α_i 及各样本的频数 f_i 键入到相应的位置（最多允许输入 30 行数值），便可获得相应的检验结果。

对于本例，将 A 病室的角度值键入（或复制粘贴）到单元格区域 B13:B18，频数键入到单元格区域 E13:E18。因为是非频数数据，所以频数的键入均为 1 即可；以同样方式，将 B 病室的角度键入到 H13:H19，键入频数“1”到 K13:K19。由 $r=0.9855$ 获得校正因子为 1.0150（查阅光盘文件“12-14 多样本比较的圆形分布分析.xlsx”中的“校正因子 K”工作表），修改单元格 H5 的值。此后便可获得图 12-36 的结果。

由图 12-36 可见，两样本合并集中测度 $r=0.9855>0.45$。经 F 检验得到 F=10.3664，分子、分母自由度分别为 1、11，对应的 P 值为 0.0082，表示两病室病人入睡时间的早晚有差别。

采用光盘文件“12-13 圆形分布分析.xlsx”，将上述相应数据复制粘贴到相应位置，可获得 A 病室的平均角 $\bar{\alpha}_1=319.59$，即 21 点 18 分 21 秒；B 病室的平均角 $\bar{\alpha}_2=333.22$，

即 22 点 12 分 52 秒。即 B 病室的平均入睡时间较 A 病室晚。



圆形分布分析

$F=K\dfrac{(N-k)\left(\sum_{i=1}^{k}R_i+R\right)}{(k-1)\left(N-\sum_{i=1}^{k}R_i\right)}$　$\nu_1=k-1,\quad \nu_2=N-k$

$\bar{y}=\frac{\sum f_i\times y_i}{n}$	$\bar{x}=\frac{\sum f_i\times x_i}{n}$	合并的集中测度 $r=\sqrt{\bar{x}^2+\bar{y}^2}$	N	k	ΣR_i	K	F	ν_1	ν_2	P
-0.5377	0.8259	0.9855	13	2	12.90242	1.0150	10.3664	1	11	0.0082

	R	R_1	R_2	R_3
	12.81181921	5.951	6.951	

	$\bar{y}_1=\frac{\sum f_i\times y_i}{n}$	$\bar{x}_1=\frac{\sum f_i\times x_i}{n}$	$r_1=\sqrt{\bar{x}_1^2+\bar{y}_1^2}$
	-0.6430	0.7552	0.9919

n_1= 6

序号 i	角度（或组中值） α_i	$y_i=\sin\left(\alpha_i\times\frac{\pi}{180}\right)$	$x_i=\cos\left(\alpha_i\times\frac{\pi}{180}\right)$	频数 f_i
1	307.5	-0.79335	0.608761429	1
2	315	-0.70711	0.707106781	1
3	318.75	-0.65935	0.751839807	1
4	320	-0.64279	0.766044443	1
5	326.25	-0.55557	0.831469612	1
6	330	-0.5	0.866025404	1

n_2= 7

序号 i	角度（或组中值） α_i	$y_i=\sin\left(\alpha_i\times\frac{\pi}{180}\right)$	$x_i=\cos\left(\alpha_i\times\frac{\pi}{180}\right)$	频数 f_i
1	322.5	-0.608761429	0.7933533	1
2	326.25	-0.555570233	0.8314696	1
3	331.25	-0.480988769	0.8767268	1
4	333.75	-0.44228869	0.8968727	1
5	335	-0.422618262	0.9063078	1
6	341.25	-0.321439465	0.9469301	1

序号 i	角度（或组中值） α_i	$\sin\left(\alpha_i\times\frac{\pi}{180}\right)$
1		

图 12-36　两病室入睡时间的比较

例 12-11：某医生检测 7～9 岁与 10～13 岁儿童心电轴的位置频数分布数据如表 12-7 所示（摘自郭祖超主编. 医用数理统计方法（第三版），1988:129），问两组儿童的心电轴有无差异。

表 12-7　两组儿童的心电轴

心电轴(°)	人　数	
	7～9 岁组	10～13 岁组
-55	1	1
-10	1	3
10	2	9
30	13	29
50	31	55
70	53	99
90	24	43
110	4	0
130	0	2
225	0	1
合 计	129	242

解：在光盘文件“12-14 多样本比较的圆形分布分析.xlsx”中，工作表“多样本比较 1”

与“多样本比较 2”的程序是一样的，只是为了读者方便保留了两个多样本比较的数据。工作表“多样本比较 2”对应于本例。对于本例，先将 11 个心电轴角度值键入到单元格区域 B13:B22，然后将这些值复制粘贴至单元格区域 H13:H22 中；将 7～9 岁组频数键入到单元格区域 E13:E22，同时将 10～13 岁组频数键入到单元格区域 K13:K22 中。由合并集中测度 $r=0.9138$ 获得校正因子为 1.0616，修改单元格 H5 的值。此后便可获得如图 12-37 所示的结果。

圆形分布分析

$\bar{y}=\frac{\sum f_i\times y_i}{n}$	$\bar{x}=\frac{\sum f_i\times x_i}{n}$	合并的集中测度 $r=\sqrt{\bar{x}^2+\bar{y}^2}$	N	k	ΣR_i	K	$F=K\frac{(N-k)\left(\sum_{i=1}^{k}R_i-R\right)}{(k-1)\left(N-\sum_{i=1}^{k}R_i\right)}$	$\nu_1=k-1,\ \nu_2=N-k$		
0.8110	0.4210	0.9138	371	2	339.092	1.0616	1.0853	1	369	0.2982
		R	339.003166							
		R_1	118.885				R_2	220.207		

R_3

$\bar{y}_1=\frac{\sum f_i\times y_i}{n}$	$\bar{x}_1=\frac{\sum f_i\times x_i}{n}$	$r_1=\sqrt{\bar{x}_1^2+\bar{y}_1^2}$								
0.8307	0.3990	0.9216				0.8005	0.4328	0.9099		

	角度		n_1= 129				角度		n_2= 242				角度		n_3= 0		
序号	(或组中值)	y_i	x_i	频数		序号	(或组中值)	y_i	x_i	频数		序号	或组中值)	y_i	x_i	频数	
i	α_i	$\sin\left(\alpha_i\times\frac{\pi}{180}\right)$	$\cos\left(\alpha_i\times\frac{\pi}{180}\right)$	f_i		i	α_i	$\sin\left(\alpha_i\times\frac{\pi}{180}\right)$	$\cos\left(\alpha_i\times\frac{\pi}{180}\right)$	f_i		i	α_i	$\sin\left(\alpha_i\times\frac{\pi}{180}\right)$	$\cos\left(\alpha_i\times\frac{\pi}{180}\right)$	f_i	
1	-55	-0.81915	0.57357644	1		1	-55	-0.819152044	0.5735764	1		1					
2	-10	-0.17365	0.98480775	1		2	-10	-0.173648178	0.9848078	3							
3	10	0.173648	0.98480775	2		3	10	0.173648178	0.9848078	9							
4	30	0.5	0.8660254	13		4	30	0.5	0.8660254	29							
5	50	0.766044	0.64278761	31		5	50	0.766044443	0.6427876	55							
6	70	0.939693	0.34202014	53		6	70	0.939692621	0.3420201	99							
7	90	1	6.1257E-17	24		7	90	1	6.126E-17	43							
8	110	0.939693	-0.3420201	4		8	110	0.939692621	-0.34202	0							
9	130	0.766044	-0.6427876	0		9	130	0.766044443	-0.642788	2							
10	225	-0.70711	-0.7071068	0		10	225	-0.707106781	-0.707107	1							

图 12-37 两年龄组心电轴的比较

由图 12-37 可见，两样本合并集中测度 $r=0.9138>0.45$。经 F 检验得到 F=1.0853，分子、分母自由度分别为 1、369，对应的 P 值为 0.2982，表示两年龄组心电轴的差别无统计学意义。

采用光盘文件“12-13 圆形分布分析.xlsx”，将上述相应数据复制粘贴到相应位置，可获得 7～9 岁与 10～13 岁的平均角分别为 $\bar{\alpha}_1=64.34$ 与 $\bar{\alpha}_2=61.60$，平均角差异不明显。

温馨提示：光盘文件“12-14 多样本比较的圆形分布分析.xlsx”中有 4 个工作表，工作表“多样本比较 1”、“多样本比较 2”用于多个平均角的比较，工作表“校正因子 K”用于查阅“校正因子”K 值，工作表“角度变换”用于时间、月份的角度变换。一般情况下，只需键入角度、频数、以及“校正因子”，便可完成多个平均角的比较。

第 13 章

寿　命　表

寿命表（life table）是根据某地特定人群的年龄组死亡率编制而成的。一般可分为队列寿命表(cohort life table)和现时寿命表(current life table)。由于人的生命周期很长，用队列寿命表的方法去研究人群的生命过程，不仅随访人数要很多，而且随访时间至少要数十年。因此，在编制寿命表时，一般不使用队列寿命表，而是使用现时寿命表。

现时寿命表是假定同时出生的一代人(一般为 10 万),按照某年(或某一时期内）特定人群的年龄组死亡率先后死去，直至死完为止，用寿命表方法计算出这一代人在不同年龄组的“死亡概率”、“死亡人数”、刚满某年龄时的“尚存人数”及其“期望寿命”等指标。这些指标不受人群年龄构成的影响，不同的地区可以相互比较。

13.1 现时寿命表

根据年龄分组方法的不同，现时寿命表可分为完全寿命表（complete life table）及简略寿命表（abridged life table）两种。完全寿命表年龄分组的组距是 1 岁，而简略寿命表习惯上将年龄组组距定为 5 岁。因为婴儿死亡率对寿命表的影响相当大，所以简略寿命表将第 1 个 5 岁年龄组拆分为组距为 1 岁的“0～”岁组，和组距为 4 岁的“1～”岁组，从 5 岁开始年龄组组距为 5 岁。这里所指年龄为实足年龄（exact age）。每一年龄组的下限值记为 X。下面介绍几个常用寿命表的概念。

1. 平均存活年数

在“X～$X+i$”岁年龄组（i 为组距），每位死亡者平均存活年数（average number of years lived）记为 a_X。完全寿命表每一年龄组平均存活年数设为 0.5 年，简略寿命表组距为 5 的年龄组平均存活年数设为 2.5 年。但由于婴儿或幼儿的早期死亡率较高，特别是新生儿第一周死亡占有较大的比重，因此寿命表的“0～”岁组平均存活年数远小于 0.5。由我国 1981 年与 1982 年部分地区的婴儿死亡资料计算得到“0～”岁组男性平均存活年数为 a_0=0.1450，女性为 a_0=0.1525，取 2 位小数，则男女统一用 a_0=0.15。当期望寿命为 75 岁时，有研究表明 0～85 岁的年龄组平均存活年数为“0.114、1.578、2.5、2.5、2.687、2.629、2.565、2.594、2.646、2.671、2.69、2.699、2.691、2.673、2.654、2.625、2.593、2.523、5.956”。

2. 年龄组死亡率

某年龄组的人口数（number of persons in the age group）记为 P_X，相应的实际死亡人数（number of deaths in the age group）记为 D_X。两者之比：

$$m_X = \frac{D_X}{P_X} \tag{13-1}$$

就是某年龄组死亡率（mortality rate）。

3. 年龄组死亡概率

“X～$X+i$”岁年龄组的死亡概率（probability of dying）记为 q_X，其计算基于年龄组死亡率 m_X，表示 X 岁尚存者在今后 i 年内死亡的概率。该指标与死亡率 m_X的意义完全不同，年龄组死亡率只是表示某年龄组人口的平均死亡水平，但这两个指标间存在下列函数关系：

$$q_x = \frac{i \times m_X}{1 + (i - a_X) \times m_X} \tag{13-2}$$

此外，还有下列特殊情况：“0～”组死亡概率一般以婴儿死亡率代替；最后一个年龄组通常是大于等于某年龄（如≥85 岁）的死亡人数，其平均存活年数 a_X 无法知晓，所以不能采用公式（13-2）计算。但因为所假定的个体一定在此年龄组死亡，所以最后一个年

龄组的死亡概率为 1。

生存概率（probability of survival）表示 X 岁尚存者在今后 i 年内存活的概率，记为 P_X，P_X=1−q_x。

4. 尚存人数

尚存人数（number of survivors）记为 l_X，表示假想的同时出生的一代人中，X 岁尚存者的平均人数。一般假定“0～”岁组的人数为 l_0=100000。

5. 死亡人数

这里是指“理论”死亡人数（number of dying），记为 d_X。表示假想的同时出生的一代人中，X 岁尚存者按死亡概率 q_X 死于年龄组“X～$X+i$”的平均人数，即

$$d_X = q_x \times l_X \tag{13-3}$$

6. 生存人年数

生存人年数（number of survival person-years）记为 L_X，是指假想的同时出生的一代人中，X 岁尚存者在今后 i 年内的平均生存人年数。计算公式为：

$$\begin{aligned} L_X &= \text{期内活满}i\text{年者人年数} + \text{期内死亡者人年数} = i \times l_{X+i} + \frac{i}{2}(l_X - l_{X+i}) \\ &= \frac{i}{2}(l_X + l_{X+i}) \end{aligned} \tag{13-4}$$

但“0～”岁组的平均生存人年数计算公式为：

$$\begin{aligned} L_0 &= \text{活满一周岁人数} + \text{期内死亡者平均存活年数} \times \text{期内死亡人数} \\ &= l_1 + a_0 \times d_0 \end{aligned} \tag{13-5}$$

最后一个年龄组死亡率、尚存人数、“理论”死亡人数和生存人年数分别记为 m_w、l_w、d_w 和 L_w。鉴于该年龄组有

$$m_w = \frac{d_w}{L_w}, \quad d_w = l_w$$

则最后一组生存人年数计算公式为：

$$L_w = \frac{l_w}{m_w} \tag{13-6}$$

7. 生存总人年数

生存总人年数（total number of survival person-years）记为 T_X，是指假想的同时出生的一代人中，X 岁尚存者今后存活的平均总人年数，它是 X 岁及 X 岁以上的各年龄组平均生存人年数的总和，即

$$T_X = \sum L_X = T_{X+i} + L_X \tag{13-7}$$

由公式可见，计算应从最大年龄组开始累加平均生存人年数。最大年龄组的 $T_X = L_X$。

8. 期望寿命

期望寿命（life expectancy）记为 e_X，这是寿命表最广泛使用的指标，是指同时出生的一代人活到 X 岁时，尚能生存的平均年数：

$$e_X = \frac{T_X}{l_X} \tag{13-8}$$

给定某一时点所有年龄组的死亡率，由寿命表计算所得出生时的期望寿命（e_0）为假想的同时出生的一代人未来的平均存活年数。这是不同国家、不同时期死亡水平比较的最常用指标。

一般情况下，随着 X 的增大，期望寿命会减少，但在较高婴儿死亡率的国家，“1～”岁组期望寿命可能反而会高于“0～”岁组期望寿命。

可以编制全国或某地区、城市或农村的简略寿命表。因为不同性别年龄组死亡率不同，故寿命表应分性别编制。下面以实例具体说明采用 Excel 编制现时简略寿命表的方法。

例 13-1：已知我国某城市 2006 年女性的年平均人口数和实际死亡人数（见图 13-1 第 F、G 列），请用该数据编制现时简略寿命表。

	A	B	D	E	F	G	H	I	J	K	L	M	N
1	简略寿命表												
2	年龄组（岁）		组距	死者平均存活年数	年平均人口数	实际死亡人数	死亡率	死亡概率	尚存人数	死亡人数	生存人年数	生存总人年数	期望寿命
3	X～	X	i	a_X	P_X	D_X	m_X	q_X	l_X	d_X	L_X	T_X	e_X
4	0～	0	1	0.15	20005	156	0.007798	0.007798	100000	780	99337	7566559	75.67
5	1～	1	4	2	86920	58	0.000667	0.002666	99220	264	396352	7467221	75.26
6	5～	5	5	2.5	102502	47	0.000459	0.002290	98956	227	494212	7070870	71.45
7	10～	10	5	2.5	151494	49	0.000323	0.001616	98729	160	493247	6576657	66.61
8	15～	15	5	2.5	182932	48	0.000262	0.001311	98570	129	492525	6083411	61.72
9	20～	20	5	2.5	203107	68	0.000335	0.001673	98440	165	491790	5590886	56.79
10	25～	25	5	2.5	240289	120	0.000499	0.002494	98276	245	490766	5099096	51.89
11	30～	30	5	2.5	247076	162	0.000656	0.003273	98031	321	489351	4608330	47.01
12	35～	35	5	2.5	199665	160	0.000801	0.003999	97710	391	487572	4118979	42.16
13	40～	40	5	2.5	163820	294	0.001795	0.008933	97319	869	484422	3631407	37.31
14	45～	45	5	2.5	145382	417	0.002868	0.014239	96450	1373	478815	3146986	32.63
15	50～	50	5	2.5	86789	522	0.006015	0.029627	95076	2817	468339	2668171	28.06
16	55～	55	5	2.5	69368	546	0.007871	0.038596	92259	3561	452395	2199832	23.84
17	60～	60	5	2.5	51207	628	0.012264	0.059496	88699	5277	430300	1747437	19.70
18	65～	65	5	2.5	39112	891	0.022781	0.107766	83421	8990	394632	1317137	15.79
19	70～	70	5	2.5	20509	831	0.040519	0.183959	74431	13692	337926	922505	12.39
20	75～	75	5	2.5	12301	926	0.075278	0.316776	60739	19241	255593	584579	9.62
21	80～	80	5	2.5	6586	731	0.110993	0.434421	41498	18028	162422	328985	7.93
22	85～	85			1909	269	0.140911	1.000000	23471	23471	166563	166563	7.10

图 13-1 某地的 2006 年女性的现时简略寿命表

解： Excel 编制简略寿命表按如下步骤进行。

① 按图 13-1 所示（也见光盘文件“13-1 简略寿命表.xlsx”）输入第 1、2、3 行文字及符号（这里的第 3 行主要由公式编辑器获得）。

② 在第 F、G 列输入各年龄组年平均人口数和实际死亡人数。这是建立寿命表的基本数据。此外，输入第 E 列数字。第 E 列是年龄组死亡者平均存活年数，假定死亡的分布是均匀的，则 4 岁年龄组为 2，5 岁年龄组为 2.5。0 岁组在刚出生时死亡较多，后期较少，按以往调查年死亡者平均存活年数只有 0.15（也见光盘文件“13-1 简略寿命表.xlsx”的第

E 列）。

③产生年龄组下限值。在单元格 B4、B5、B6 中分别键入 0、1、5。因为此后的每组年龄组组距均为 5，所以如果该行年平均人口数和死亡数所在单元格有数据，则该组的年龄下限值为“上一组年龄+5”，如单元格 B7 中的 Excel 公式为“=IF(AND(F7<>"",G7<>""),B6+5,"")”，将此公式向下拖放至单元格 B25。

④产生组距 i。在单元格 D4 中键入公式“=IF(B4<>"",IF(B5<>"",B5-B4,""),"")”，将此公式向下拖放至单元格 D25。因为最后一组的组距是不确定的，所以到最后一组不显示组距值。

⑤产生各年龄组名称。由年龄组下限值（第 B 列）和符号“～”（第 C 列）获得，即在单元格 A4 中键入公式“=IF(AND(F4<>"",G4<>""),B4&C4,"")”，将此公式向下拖放至单元格 A25。

⑥用各年龄组死亡人数除以年平均人口数获得第 H 列的年龄组死亡率（m_X）。先单击单元格 H4 中输入公式“=IF(AND(F4<>"",G4<>""),G4/F4,"")”，将此公式向下拖放至单元格 H25。

⑦计算年龄组死亡概率（q_X），0 岁组等于婴儿死亡率，即单元格 I4 为“=H4”。最后一组为 1，其他组由公式（13-2）计算。在单元格 I5 中键入公式“=IF(AND(F5<>"",G5<>""),IF(E5<>"",D5*H5/(1+(D5-E5)*H5),1),"")”，将此公式向下拖放至单元格 I25。

⑧计算尚存人数与死亡人数。在单元格 K4 中键入公式“=IF(AND(F4<>"",G4<>""),I4*J4,"")”，由公式（13-3）计算死亡人数。在单元格 J4 中键入 100000；在单元格 J5 中键入公式“=IF(AND(F5<>"",G5<>""),J4-K4,"")”，并将此公式拖放至单元格 J25，求尚存人数。

⑨在第 L 列计算生存人年数。采用公式（13-5），在单元格 L4 中键入“=J5+K4*E4”，获得 0 岁生存人年。采用公式（13-4）和公式（13-6），在单元格 L5 中键入公式“=IF(AND(F5<>"",G5<>""),IF(E5<>"",D5*(J5+J6)/2,J5/H5),"")”，并将此公式拖放至单元格 L25，求其他组的生存人年数。

⑩在第 M、N 列分别计算生存总人年数和期望寿命。先在单元格 M4 中键入公式“=IF(AND(F4<>"",G4<>""),IF(E4<>"",M5+L4,L4),"")”，然后在单元格 N4 中键入公式“=IF(AND(F4<>"",G4<>""),M4/J4,"")”，选取单元格区域 M4:N4，将此区域拖放至 N25。

结果分析：由图 13-1 可见，该地 2006 年女性出生期望寿命为 75.67。其他每个年龄组的期望寿命见图 13-1 的最后一列所示。在光盘文件“13-1 简略寿命表.xlsx”右侧（如图 13-2 所示），程序根据前面的计算经过，自动绘制了 4 个图形，分别反映随年龄变化，死亡概率、尚存人数、死亡人数、期望寿命的变化情况。注意死亡概率与期望寿命对应的“年龄”是具体数字，即各年龄组下限值绘制（见程序第 B 列），而尚存人数与死亡人数对应

的“年龄”是分类标志（见程序第 A 列）。但应注意，尚存人数与死亡人数对应图形中“0～”与“1～”岁年龄组不等距，看图是只能将“年龄”作为分类标志看待，否则容易误解。通过比较不同地区、不同国家相同性别的死亡概率、尚存人数、死亡人数、期望寿命等指标随年龄的变化情况，可以反映当地的社会卫生状况。

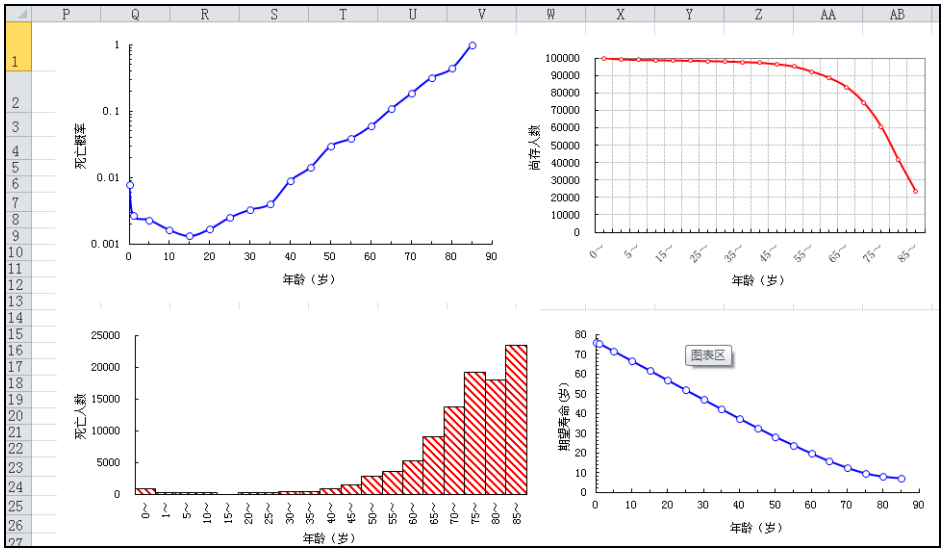


图 13-2 简略寿命表图示

有时无“年平均人口数”和“实际死亡人数”的具体数据，只有由这两个指标计算获得的年龄组“死亡率”。此时，可采用光盘文件“13-1 简略寿命表.xlsx”中的“简略寿命表（2）”工作表。将年龄组“死亡率”数据键入到该工作表的单元格区域 F4:F25 中，并注意适当修改单元格 E4 的“0～”岁组平均存活年数，便可同样获得简略寿命表数据结果。比如由例 13-1 数据中的“死亡率”数据获得的结果如图 13-3 所示（可见光盘文件“13-1 简略寿命表.xlsx”的“简略寿命表（2）”工作表）。当期望寿命为 75 岁时，有研究表明 0～85 岁的年龄组平均存活年数 a_X 为“0.114、1.578、2.5、2.5、2.687、2.629、2.565、2.594、2.646、2.671、2.69、2.699、2.691、2.673、2.654、2.625、2.593、2.523、5.956”，由此获得期望寿命变化表达（见光盘文件“13-1 简略寿命表.xlsx”的“简略寿命表（3）”工作表）。

温馨提示：程序设置可以建立 100 岁以下的简略寿命表，可以自动判断哪一组属于最后一个年龄组。使用时只需将已知的各年龄组“年平均人口数”和“实际死亡人数”分别依次键入到单元格区域 F4:F25、G4:G25，并注意适当修改单元格 E4 的“0～”岁组平均存活年数即可。最后一个年龄组可以是“50～”、……、“75～”、……、“100～”，程序可自动判断。

简略寿命表（已知死亡率）

年龄组（岁）		组距	死者平均存活年数	死亡率	死亡概率	尚存人数	死亡人数	生存人年数	生存总人年数	期望寿命
X~	X	i	a_X	m_X	q_X	l_X	d_X	L_X	T_X	e_X
0~	0	1	0.15	0.007798	0.007798	100000	780	99337	7566559	75.67
1~	1	4	2	0.000667	0.002666	99220	264	396352	7467221	75.26
5~	5	5	2.5	0.000459	0.002290	98956	227	494212	7070870	71.45
10~	10	5	2.5	0.000323	0.001616	98729	160	493247	6576657	66.61
15~	15	5	2.5	0.000262	0.001311	98570	129	492525	6083411	61.72
20~	20	5	2.5	0.000335	0.001673	98440	165	491790	5590886	56.79
25~	25	5	2.5	0.000499	0.002494	98276	245	490766	5099096	51.89
30~	30	5	2.5	0.000656	0.003273	98031	321	489351	4608330	47.01
35~	35	5	2.5	0.000801	0.003999	97710	391	487572	4118979	42.16
40~	40	5	2.5	0.001795	0.008933	97319	869	484422	3631407	37.31
45~	45	5	2.5	0.002868	0.014239	96450	1373	478815	3146986	32.63
50~	50	5	2.5	0.006015	0.029627	95076	2817	468339	2668171	28.06
55~	55	5	2.5	0.007871	0.038596	92259	3561	452395	2199832	23.84
60~	60	5	2.5	0.012264	0.059496	88699	5277	430300	1747437	19.70
65~	65	5	2.5	0.022781	0.107766	83421	8990	394632	1317137	15.79
70~	70	5	2.5	0.040519	0.183959	74431	13692	337926	922505	12.39
75~	75	5	2.5	0.075278	0.316776	60739	19241	255593	584579	9.62
80~	80	5	2.5	0.110993	0.434421	41498	18028	162422	328985	7.93
85~	85			0.140911	1.000000	23471	23471	166563	166563	7.10

图 13-3　由“死亡率”编制的简略现时寿命表

如果需要编制完全寿命表（编制原理与前面简略寿命表十分相似），可以打开光盘文件“13-2 完全寿命表.xlsx”的“完全寿命表”工作表，如图 13-4 所示。使用方法与简略寿命表相同，在此不再赘述。

完全寿命表

年龄组（岁）		组距	死者平均存活年数	年平均人口数	实际死亡人数	死亡率	死亡概率	尚存人数	死亡人数	生存人年数	生存总人年数	期望寿命
X~	X	i	a_X	P_X	D_X	m_X	q_X	l_X	d_X	L_X	T_X	e_X
0~	0	1	0.15	96211	4691	0.048757	0.048757	100000	4876	95856	7066744	70.67
1~	1	1	0.5	94971	674	0.007097	0.007072	95124	673	94788	6970889	73.28
2~	2	1	0.5	94386	496	0.005255	0.005241	94452	495	94204	6876101	72.80
3~	3	1	0.5	93936	405	0.004311	0.004302	93957	404	93754	6781897	72.18
4~	4	1	0.5	93609	249	0.002660	0.002656	93552	249	93428	6688142	71.49
5~	5	1	0.5	93379	212	0.002270	0.002268	93304	212	93198	6594714	70.68
6~	6	1	0.5	93176	194	0.002082	0.002080	93092	194	92995	6501516	69.84
7~	7	1	0.5	93015	127	0.001365	0.001364	92899	127	92835	6408521	68.98
8~	8	1	0.5	92882	139	0.001497	0.001495	92772	139	92702	6315686	68.08
9~	9	1	0.5	92766	95	0.001024	0.001024	92633	95	92586	6222983	67.18
10~	10	1	0.5	92632	172	0.001857	0.001855	92538	172	92452	6130398	66.25
11~	11	1	0.5	92499	95	0.001027	0.001027	92367	95	92319	6037945	65.37
12~	12	1	0.5	92399	107	0.001158	0.001157	92272	107	92218	5945626	64.44
13~	13	1	0.5	92301	88	0.000953	0.000953	92165	88	92121	5853408	63.51
14~	14	1	0.5	92194	127	0.001378	0.001377	92077	127	92014	5761287	62.57
15~	15	1	0.5	92070	121	0.001314	0.001313	91950	121	91890	5669273	61.66
16~	16	1	0.5	91941	138	0.001501	0.001500	91830	138	91761	5577383	60.74
17~	17	1	0.5	91788	166	0.001809	0.001807	91692	166	91609	5485622	59.83
18~	18	1	0.5	91633	145	0.001582	0.001581	91526	145	91454	5394013	58.93
19~	19	1	0.5	91479	163	0.001782	0.001780	91382	163	91300	5302559	58.03

图 13-4　完全寿命表

13.2　去死因寿命表

研究某种死因对居民死亡的影响，可编制去死因寿命表（cause eliminated life table）。

其基本思想是：假使消除了某种死因，则原死于该原因的人不死于该原因，寿命就会有所延长。显然，如果消除了对生命威胁大的死因，寿命就会延长更多。去死因寿命表的优点是：

①以某死因耗损的期望寿命和尚存人数合理地说明了该死因对人群生命的影响程度。

②去死因寿命表的指标既能综合说明某死因对全人口的作用，又能分别说明某死因对各年龄组人口的作用。

③去死因寿命表的指标同样不受人口年龄构成的影响，便于相互比较。

例 13-2：已知某地的 2006 年女性的各年龄组年平均人口数、全死因死亡数及肿瘤死亡数（如图 13-5 第 F～H 列所示），请用这些数据求各年龄组尚存者去肿瘤死亡后的期望寿命。

解：去死因寿命表中各项指标的意义与全死因寿命表（即“13.1 现时寿命表”）相同，为了区分，在有关符号的右上角加标记“′”，表示某死因或去某死因。

编制去死因寿命表的关键是求去某死因后各年龄组的生存概率（p'_X），有了 p'_X，就可仿照前面介绍的编制全死因寿命表的方法，来编制去某死因寿命表。步骤如下：

①按图 13-5（也见光盘文件“13-3 去死因寿命表.xlsx”）输入第 1、2、3、4 行文字及符号，图中的第 A、B、C、D、E 列与图 13-1 相同（只是数据下移一行），第 H 列的某死因死亡人数 D'_X 是编制去死因寿命表必须具备的基本数据。

	A	B	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
2											去某死因死亡后							
3	年龄组（岁）		组距	死者平均存活年数	年平均人口数	实际死亡人数	某死因死亡人数	去某死因死亡比例	死亡概率	生存概率	生存概率	尚存人数	死亡人数	生存人年数	生存总人年数	预期寿命	该死因损失的寿命	损失百分率
4	X～	X	i	a_x	P_X	D_X	D'_X	r'_X	q_x	p_x	p'_X	l'_X	d'_x	L'_X	T'_X	e'_X	e'_x-e_x	$(e'_x-e_x)/e_x$
5	0～	0	1	0.15	20005	156	2	0.987179	0.007798	0.992202	0.992302	100000	770	99346	7903987	79.04	3.37	4.46%
6	1～	1	4	2	86920	58	4	0.931034	0.002666	0.997334	0.997518	99230	246	396428	7804642	78.65	3.39	4.51%
7	5～	5	5	2.5	102502	47	8	0.829787	0.002290	0.997710	0.998099	98984	188	494449	7408213	74.84	3.39	4.74%
8	10～	10	5	2.5	151494	49	11	0.775510	0.001616	0.998384	0.998747	98796	124	493669	6913764	69.98	3.37	5.05%
9	15～	15	5	2.5	182932	48	13	0.729167	0.001311	0.998689	0.999044	98672	94	493124	6420095	65.07	3.35	5.43%
10	20～	20	5	2.5	203107	68	21	0.691176	0.001673	0.998327	0.998844	98578	114	492603	5926972	60.12	3.33	5.86%
11	25～	25	5	2.5	240289	120	36	0.700000	0.002494	0.997506	0.998254	98464	172	491888	5434369	55.19	3.31	6.37%
12	30～	30	5	2.5	247076	162	41	0.746914	0.003273	0.996727	0.997554	98292	240	490857	4942481	50.28	3.27	6.97%
13	35～	35	5	2.5	199665	160	44	0.725000	0.003999	0.996001	0.997099	98051	284	489545	4451624	45.40	3.25	7.70%
14	40～	40	5	2.5	163820	294	80	0.727891	0.008933	0.991067	0.993490	97767	636	487243	3962079	40.53	3.21	8.61%
15	45～	45	5	2.5	145382	417	142	0.659472	0.014239	0.985761	0.990587	97130	914	483366	3474836	35.77	3.15	9.64%
16	50～	50	5	2.5	86789	522	210	0.597701	0.029627	0.970373	0.982185	96216	1714	476795	2991470	31.09	3.03	10.79%
17	55～	55	5	2.5	69368	546	315	0.423077	0.038596	0.961404	0.983485	94502	1561	468608	2514675	26.61	2.77	11.60%
18	60～	60	5	2.5	51207	628	360	0.426752	0.059496	0.940504	0.974163	92941	2401	458703	2046068	22.01	2.31	11.74%
19	65～	65	5	2.5	39112	891	381	0.572391	0.107766	0.892234	0.936816	90540	5721	438398	1587365	17.53	1.74	11.04%
20	70～	70	5	2.5	20509	831	248	0.701564	0.183959	0.816041	0.867082	84819	11274	395911	1148967	13.55	1.15	9.30%
21	75～	75	5	2.5	12301	926	127	0.862851	0.316776	0.683224	0.719867	73545	20602	316220	753056	10.24	0.61	6.39%
22	80～	80	5	2.5	6586	731	60	0.917921	0.434421	0.565579	0.592664	52943	21566	210800	436835	8.25	0.32	4.08%
23	85～	85			1909	269	4	0.985130	1.000000	0.000000	0.000000	31377	31377	226035	226035	7.20	0.11	1.51%

图 13-5 某地 2006 年女性去肿瘤简略寿命表

②由第 F、G 列基本数据，根据公式：

$$r'_X = \frac{D_X - D'_X}{D_X} \tag{13-9}$$

计算去某死因后的死亡人数占全死因人数的比例 r'_X（第 I 列）。在单元格 I5 处输入 Excel 公式“=IF(AND($G5<>"",$H5<>""),(G5-H5)/G5,"")”，并将此公式向下拖放填充至 I26。

③计算全死因寿命表死亡概率 q_x（第 J 列），计算方法同现时简略寿命表编制步骤的

第⑦步。先在单元格 J5 处输入公式“=IF(AND(F5<>"",G5<>""),G5/F5,"")”，单击“V（输入）”按钮（或按 Enter 键）在单元格 J6 处输入公式“=IF(AND(F6<>"",G6<>""),IF(E6<>"",D6*(G6/F6)/ (1+(D6-E6)*(G6/F6)),1),"")”，按 Enter 键然后将此公式向下拖放至 J26，由此获得“1～”岁及以上年龄组的死亡概率。程序自动判断是否为最后一个年龄组，如果是则给出固定值“1”。

④计算全死因寿命表生存概率 p_x（第 K 列）。采用 1 减第 J 列即可。先在单元格 K5 处输入公式“=IF(J5<>"",1-J5,"")”，然后将此公式向下拖放至 K26。

⑤计算去某死因后的生存概率 p'_x（第 L 列），这是编制去死因寿命表的关键指标，计算公式为：

$$p'_x = (p_x)^{r'_x} \tag{13-10}$$

在单元格 L5 处输入公式“=IF(AND($G5<>"",$H5<>""),K5^I5,"")”，然后将此公式向下拖放填充至 L26。

⑥计算去某死因死亡后的尚存人数 l'_X（第 M 列）与死亡人数 d'_X（第 N 列）。

令 $l'_0 = 100000$，$l'_{X+i} = l'_X - d'_X$，计算去某死因死亡后的尚存人数。利用第 L、M 列，按公式：

$$d'_X = (1 - p'_x) \times l'_X \tag{13-11}$$

计算去某死因死亡后的死亡人数。

在单元格 M5 处输入 100000，在单元格 M6 处输入公式“=IF(AND($G6<>"",$H6<>""),M5-N5,"")”，然后将单元格 M6 中的公式向下拖放填充至 M26，计算尚存人数；在单元格 N5 处输入公式“=IF(AND($G5<>"",$H5<>""),(1–L5)*M5,"")”，然后将单元格 N5 中的公式向下拖放填充至 N26，获得去某死因死亡后的死亡人数。

⑦计算去某死因死亡后的生存人年数 L'_X（第 K 列）。计算方法与全死因寿命表相同。在单元格 O5 处输入公式“=M6+N5*E5”；在单元格 O6 处输入公式“=IF(H6<>"",IF(E6<>"",D6*(M6+M7)/2,M6/((G6–H6)/F6)),"")”；然后将 O6 中公式向下拖放至 O26。程序自动判断是否为最后一个年龄组，如果是则计算公式为“M6/((G6–H6)/F6))”，否则为“D6*(M6+M7)/2”。

⑧计算各年龄组去某死因死亡后的生存总人年数 T'_X（第 P 列）。从最后一组开始，对去某死因死亡后的生存人年数 L'_X 进行累加。

在单元格 P5 处输入公式“=IF(H5<>"",IF(D5<>"",P6+O5,O5),"")”，将公式向下拖放填充至单元格 P26。程序自动判断是否为最后一个年龄组，如果是则计算公式为“O5”，否则为“P6+O5”。

⑨计算去某死因死亡后的期望寿命 e'_X（第 Q 列）。由去某死因死亡后生存总人年数与尚存人数的数据之比获得，如去某死因死亡后的出生时期望寿命 $e'_0 = 7903987/100000 = 79.04$。

在单元格 Q5 处输入公式“=IF(H5<>"",P5/M5,"")”，将该公式向下拖放填充至 Q26 即可。

⑩ 第 R、S 列分别为该死因损失的寿命、该死因寿命损失百分率。这两个指标涉及前面所介绍的现时寿命表，因此在光盘文件“13-3 去死因寿命表.xlsx”中也有“简略寿命表”工作表，其数据直接应用“去某死因寿命表”工作表。也就是说，只要改变“去某死因寿命表”工作表数据，“简略寿命表”工作表数据也将自动改变。

结果分析：该地 2006 年女性因某肿瘤死亡而减少的出生时期望寿命为：

$$e_0' - e_0 = 79.03 - 75.07 = 3.37 \text{（岁）}$$

也就是说，由于肿瘤损失的出生时期望寿命为 3.37 岁，如果某死因的死亡得到了有效的控制，出生时期望寿命可增加 $(e_0' - e_0)/e_0 = (79.03 - 75.07)/75.07 = 4.46\%$。

其他年龄组的寿命损失年数及寿命损失百分率如图 13-5 的第 R、S 列所示（见光盘文件“13-3 去死因寿命表.xlsx”的“去某死因寿命表”工作表）。

温馨提示：程序设置可以建立 100 岁以下的去某死因寿命表，并可以自动判断哪一组属于最后一个年龄组。使用时只需将已知的各年龄组“年平均人口数”、“实际死亡人数”和“某死因死亡人数”分别依次键入到单元格区域 F5:F26、G5:G26 和 H5:H26，并注意适当修改单元格 E5 的“0～”岁组平均存活年数即可。最后一个年龄组可以是“50～”、……、“75～”、……、“100～”，程序可自动判断。

如果需要编制去死因完全寿命表（编制原理与前面的“去死因简略寿命表”十分相似），可以打开光盘文件“13-4 去死因完全寿命表.xlsx”的“去某死因完全寿命表”工作表，如图 13-6 所示。使用方法与上述介绍的“去死因简略寿命表”相同，在此不再赘述。

	A	B	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	去某死因完全寿命表																	
2											去某死因死亡后							
3	年龄组（岁）		组距	死者平均存活年数	年平均人口数	实际死亡人数	某死因死亡人数	去某死因死亡比例	死亡概率	生存概率	生存概率	尚存人数	死亡人数	生存人年数	生存总人年数	预期寿命	该死因损失的寿命	损失百分率
4	X～	X	i	a_x	P_x	D_x	D_x'	r_x'	q_x	p_x	p_x'	l_x'	d_x'	L_x'	T_x'	e_x'	$e_x'-e_x$	$\lvert e_x'-e_x\rvert/e_x$
5	0～	0	1	0.15	96211	4691	2	0.999574	0.048757	0.951243	0.951263	100000	4874	95857	7133955	71.34	0.67	0.95%
6	1～	1	1	0.5	94971	674	4	0.994065	0.007072	0.992928	0.992970	95126	669	94792	7038097	73.99	0.70	0.96%
7	2～	2	1	0.5	94386	496	8	0.983871	0.005241	0.994759	0.994843	94458	487	94214	6943305	73.51	0.71	0.97%
8	3～	3	1	0.5	93936	405	11	0.972840	0.004302	0.995698	0.995814	93970	393	93774	6849091	72.89	0.70	0.98%
9	4～	4	1	0.5	93609	249	13	0.947791	0.002656	0.997344	0.997482	93577	236	93459	6755318	72.19	0.70	0.98%
10	5～	5	1	0.5	93379	212	21	0.900943	0.002268	0.997732	0.997957	93341	191	93246	6661858	71.37	0.69	0.98%
11	6～	6	1	0.5	93176	194	36	0.814433	0.002080	0.997920	0.998306	93151	158	93072	6568612	70.52	0.68	0.97%
12	7～	7	1	0.5	93015	127	41	0.677165	0.001364	0.998636	0.999076	92993	86	92950	6475540	69.63	0.65	0.94%
13	8～	8	1	0.5	92882	139	44	0.683453	0.001495	0.998505	0.998978	92907	95	92860	6382590	68.70	0.62	0.91%
14	9～	9	1	0.5	92766	95	80	0.157895	0.001024	0.998976	0.999838	92812	15	92805	6289731	67.77	0.59	0.88%
15	10～	10	1	0.5	92632	172	21	0.877907	0.001855	0.998145	0.998371	92797	151	92721	6196926	66.78	0.53	0.80%
16	11～	11	1	0.5	92499	95	2	0.978947	0.001027	0.998973	0.998995	92646	93	92599	6104205	65.89	0.52	0.79%
17	12～	12	1	0.5	92399	107	4	0.962617	0.001157	0.998843	0.998886	92553	103	92501	6011606	64.95	0.52	0.80%
18	13～	13	1	0.5	92301	88	8	0.909091	0.000953	0.999047	0.999134	92450	80	92410	5919104	64.03	0.52	0.81%
19	14～	14	1	0.5	92194	127	11	0.913386	0.001377	0.998623	0.998743	92370	116	92311	5826695	63.08	0.51	0.82%
20	15～	15	1	0.5	92070	121	13	0.892562	0.001313	0.998687	0.998828	92253	108	92199	5734383	62.16	0.50	0.82%
21	16～	16	1	0.5	91941	138	21	0.847826	0.001500	0.998500	0.998728	92145	117	92087	5642184	61.23	0.50	0.82%
22	17～	17	1	0.5	91788	166	36	0.783133	0.001807	0.998193	0.998585	92028	130	91963	5550097	60.31	0.48	0.81%
23	18～	18	1	0.5	91633	145	41	0.717241	0.001581	0.998419	0.998866	91898	104	91846	5458134	59.39	0.46	0.78%
24	19～	19	1	0.5	91479	163	44	0.730061	0.001780	0.998220	0.998700	91794	119	91734	5366289	58.46	0.43	0.75%

图 13-6 去某死因完全寿命表

13.3 健康期望寿命表

随着生活条件的改善，人们不再只是追求客观的生存与躯体寿命的延长，而更注重于躯体、精神和社会三方面的完美状态，即更注重于生存的质量。1964 年 Sanders 首次提出了健康期望寿命的概念，1971 年 Sullivan 提出了编制健康期望寿命表的方法，该方法称为 Sullivan 方法，该方法简便易行，目前应用广泛。健康期望寿命（health expectancies）是将期望寿命分成良好健康状况和较差健康状况两种情况，同时考虑年龄组死亡率与年龄组患病率，以良好健康状况下的期望寿命来反映人群健康状况。根据年龄组患病率中“患病”含义的不同，健康期望寿命可细分为无残疾期望寿命（Disability-Free Life Expectancy，DFLE）、有活力期望寿命（Active Life Expectancy，ALE）（不依赖于他人进行日常生活）、健康生活期望寿命（Healthy Life Expectancy）（自我感觉健康）等。

例 13-3：在例 13-1 的已知我国某城市 2006 年女性的年平均人口数和实际死亡人数基础上，还已知该地某疾病的患病率（如图 13-7 的第 O 列所示，也见光盘文件“13-5 健康期望寿命表.xlsx”）。试进一步计算健康期望寿命。

	A	B	D	E	F	G	H	J	L	N	O	P	Q	R	S
1	健康期望寿命表														
2	年龄组（岁		组距	死者平均存活年数	年平均人口数	实际死亡人数	死亡率	尚存人数	生存人年数	期望寿命	患病率	健康生存人年数	累积健康生存人年数	健康期望寿命	患病损失健康期望寿命
3	X～	X	i	a_X	P_X	D_X	m_X	l_X	L_X	e_X	R_X	h_X	H_X	E_X	DLE_x
4	0～	0	1	0.15	20005	156	0.007798	100000	99337	75.67	0.053	94072	6638183	66.38	9.28
5	1～	1	4	2	86920	58	0.000667	99220	396552	75.26	0.053	375345	6544111	65.96	9.30
6	5～	5	5	2.5	102502	47	0.000459	98956	494212	71.45	0.053	468019	6168765	62.34	9.12
7	10～	10	5	2.5	151494	49	0.000323	98729	493247	66.61	0.053	467105	5700747	57.74	8.87
8	15～	15	5	2.5	182932	48	0.000262	98570	492525	61.72	0.053	466421	5233642	53.10	8.62
9	20～	20	5	2.5	203107	68	0.000335	98440	491790	56.79	0.062	461299	4767221	48.43	8.37
10	25～	25	5	2.5	240289	120	0.000499	98276	490766	51.89	0.067	457884	4305922	43.81	8.07
11	30～	30	5	2.5	247076	162	0.000656	98031	489351	47.01	0.076	452160	3848038	39.25	7.76
12	35～	35	5	2.5	199665	160	0.000801	97710	487572	42.16	0.077	450029	3395877	34.75	7.40
13	40～	40	5	2.5	163820	294	0.001795	97319	484422	37.31	0.104	434042	2945849	30.27	7.04
14	45～	45	5	2.5	145382	417	0.002868	96450	478815	32.63	0.123	419921	2511807	26.04	6.59
15	50～	50	5	2.5	86789	522	0.006015	95076	468339	28.06	0.151	397620	2091886	22.00	6.06
16	55～	55	5	2.5	69368	546	0.007871	92259	452395	23.84	0.174	373678	1694266	18.36	5.48
17	60～	60	5	2.5	51207	628	0.012264	88699	430300	19.70	0.175	354997	1320588	14.89	4.81
18	65～	65	5	2.5	39112	891	0.022781	83421	394632	15.79	0.178	324388	965590	11.57	4.21
19	70～	70	5	2.5	20509	831	0.040519	74431	337926	12.39	0.236	258176	641203	8.61	3.78
20	75～	75	5	2.5	12301	926	0.075278	60739	255593	9.62	0.293	180705	383027	6.31	3.32
21	80～	80	5	2.5	6586	731	0.110993	41498	162422	7.93	0.344	106549	202323	4.88	3.05
22	85～	85			1909	269	0.140911	23471	166563	7.10	0.425	95774	95774	4.08	3.02

图 13-7　我国某城市 2006 年女性的 Sullivan 健康期望寿命表

解：图 13-7 中第 F、G、O 列为已知数据。编制 Sullivan 健康寿命表的第 A～N 列的方法与例 13-1 的编制简略寿命表完全相同。不同之处在于：

①第 O 列为基本数据，由用户输入“年龄组患病率” R_X 。

②第 P 列的健康（无依赖）生存人年数的计算公式为：

$$h_X = L_X(1 - R_X) \tag{13-12}$$

式中，L_X 为表中第 L 列的生存人年数，R_X 为表中第 O 列的患病率。在单元格 P4 中输入

公式“=IF(O4<>"",L4*(1-O4),"")”，并将此公式向下拖放至单元格 P25。

③累积健康生存人年数的计算公式为：

$$H_X = \sum h_X = H_{X+1} + h_X \tag{13-13}$$

即在单元格 Q4 中输入公式“=IF(O4<>"",IF(E4<>"",P4+Q5,P4),"")”，将此公式向下拖放至 Q25。程序将自动判断是否为最后一个年龄组，如果是最后一个年龄组，那么累积健康生存人年数等于对应年龄组的健康生存人数，否则按公式（13-13）计算。

④第 R 列的健康期望寿命为第 Q 列的 H_X 与第 J 列尚存人数 l_X 的比值，即：

$$E_X = H_X / l_X \tag{13-14}$$

在单元格 R4 中输入公式“=IF($O4<>"",Q4/J4,"")”，将此公式向下拖放至单元格 R25。

⑤利用第 N 列的期望寿命 e_X 与第 R 列的健康期望寿命 E_X 的差值，可获得由于“患病”所损失的期望寿命（dependent life expectancy,DLE），即：

$$DLE_X = e_X - E_X \tag{13-15}$$

结果分析：该地 2006 年女性因“患病”损失的出生时健康期望寿命为：

$$e_0 - E_0 = 75.67 - 66.38 = 9.28\text{（岁）}$$

温馨提示：程序设置可以建立 100 岁以下的健康期望寿命表，并可以自动判断哪一组属于最后一个年龄组。使用时只需将已知的各年龄组“年平均人口数”、“实际死亡人数”和“患病率”分别依次键入到单元格区域 F4:F25、G4:G25 和 O4:O25，并注意适当修改单元格 E5 的“0～”岁组平均存活年数即可。最后一个年龄组可以是“50～”、……、“75～”、……、“100～”，程序可自动判断。

如果只有年龄组“死亡率”和“患病率”，则可采用光盘文件“13-5 健康期望寿命表.xlsx”的“已知死亡率”工作表，如图 13-8 所示。

健康期望寿命表（已知死亡率）

年龄组（岁）	X～	X	组距 i	死者平均存活年数 a_X	死亡率 m_X	尚存人数 l_X	生存人年数 L_X	期望寿命 e_X	患病率 R_X	健康生存人年数 h_X	累积健康生存人年数 H_X	健康期望寿命 E_X	患病损失健康期望寿命 DLE_X
0～	0	1	0.15	0.007798	100000	99337	75.67	0.053	94072	6638183	66.38	9.28	
1～	1	4	2	0.000667	99220	396352	75.26	0.053	375345	6544111	65.96	9.30	
5～	5	5	2.5	0.000459	98956	494212	71.45	0.053	468019	6168765	62.34	9.12	
10～	10	5	2.5	0.000323	98729	493247	66.61	0.053	467105	5700747	57.74	8.87	
15～	15	5	2.5	0.000262	98570	492525	61.72	0.053	466421	5233642	53.10	8.62	
20～	20	5	2.5	0.000335	98440	491790	56.79	0.062	461299	4767221	48.43	8.37	
25～	25	5	2.5	0.000499	98276	490766	51.89	0.067	457884	4305922	43.81	8.07	
30～	30	5	2.5	0.000656	98031	489351	47.01	0.076	452160	3848038	39.25	7.76	
35～	35	5	2.5	0.000801	97710	487572	42.16	0.077	450029	3395877	34.75	7.40	
40～	40	5	2.5	0.001795	97319	484422	37.31	0.104	434042	2945849	30.27	7.04	
45～	45	5	2.5	0.002868	96450	478815	32.63	0.123	419921	2511807	26.04	6.59	
50～	50	5	2.5	0.006015	95076	468339	28.06	0.151	397620	2091886	22.00	6.06	
55～	55	5	2.5	0.007871	92259	452395	23.84	0.174	373678	1694266	18.36	5.48	
60～	60	5	2.5	0.012264	88699	430300	19.70	0.175	354997	1320588	14.89	4.81	
65～	65	5	2.5	0.022781	83421	394632	15.79	0.178	324388	965590	11.57	4.21	
70～	70	5	2.5	0.040519	74431	337926	12.39	0.236	258176	641203	8.61	3.78	
75～	75	5	2.5	0.075278	60739	255593	9.62	0.293	180705	383027	6.31	3.32	
80～	80	5	2.5	0.110993	41498	162422	7.93	0.344	106549	202323	4.88	3.05	
85～	85			0.140911	23471	166563	7.10	0.425	95774	95774	4.08	3.02	

图 13-8 “已知死亡率”工作表的健康期望寿命表

例 13-4：已知某地的 2006 年男性 75 岁以上老年人的年龄组死亡率和年龄组 ADL 患病率（如图 13-9 的第 F、M 列所示，也可见光盘文件“13-6 自定义年龄健康期望寿命表.xlsx”的“已知死亡率”工作表），用这两栏已知数据求各年龄的 ADL 依赖期望寿命。

解：采用方法与例 13-3 非常相似，略为不同的是起点时间不是“0～”岁年龄组，而是其他年龄组，所以计算公式忽略了“0～”岁年龄组的特殊情况。结果及其解释与上述相同。

自定义年龄组健康期望寿命表

年龄组（岁		组距	死者平均存活年数	死亡率	死亡概率	尚存人数	生存人年数	期望寿命	患病率	健康生存人年数	累积健康生存人年数	健康期望寿命	患病损失健康期望寿命
X～	X	i	a_X	m_X	q_X	l_X	LT_X	e_X	R_X	h_X	H_X	E_X	DLE_X
75～	75	1	0.5	0.036	0.035363	100000	98232	11.01	0.104	88016	806604	8.07	2.95
76～	76	1	0.5	0.04	0.039216	96464	94572	10.40	0.104	84737	718588	7.45	2.95
77～	77	1	0.5	0.044	0.043053	92681	90686	9.80	0.104	81254	633852	6.84	2.96
78～	78	1	0.5	0.049	0.047828	88691	86570	9.22	0.104	77566	552597	6.23	2.99
79～	79	1	0.5	0.056	0.054475	84449	82149	8.66	0.104	73605	475031	5.63	3.04
80～	80	1	0.5	0.062	0.060136	79848	77447	8.13	0.203	61726	401426	5.03	3.10
81～	81	1	0.5	0.071	0.068566	75047	72474	7.62	0.203	57762	339700	4.53	3.09
82～	82	1	0.5	0.08	0.076923	69901	67212	7.14	0.203	53568	281939	4.03	3.11
83～	83	1	0.5	0.089	0.085208	64524	61775	6.70	0.203	49235	228370	3.54	3.16
84～	84	1	0.5	0.098	0.093422	59026	56269	6.27	0.203	44846	179136	3.03	3.24
85～	85	1	0.5	0.112	0.106061	53512	50674	5.87	0.543	23158	134289	2.51	3.36
86～	86	1	0.5	0.126	0.118532	47836	45001	5.50	0.543	20566	111131	2.32	3.18
87～	87	1	0.5	0.134	0.125586	42166	39518	5.18	0.543	18060	90566	2.15	3.03
88～	88	1	0.5	0.148	0.137803	36871	34330	4.85	0.543	15689	72506	1.97	2.88
89～	89	1	0.5	0.169	0.155832	31790	29313	4.55	0.543	13396	56817	1.79	2.76
90～	90			0.233	1.000000	26836	115175	4.29	0.623	43421	43421	1.62	2.67

图 13-9　自定义年龄的健康期望寿命表

温馨提示：程序设置可以建立 100 岁以下的健康期望寿命表，并可以自动判断哪一组属于最后一个年龄组。使用时只需修改年龄组下限值 X（第 B 列）、年龄组“死亡率”（第 F 列）和“患病率”（第 M 列）即可。

如果有各年龄组“年平均人口数”、“实际死亡人数”和“患病率”，则可采用光盘文件“13-6 自定义年龄健康期望寿命表.xlsx”的“已知人口与死亡数”工作表，如图 13-10 所示。

自定义年龄组健康期望寿命表

组距	死者平均存活年数	年平均人口数	实际死亡人数	死亡率	尚存人数	生存人年数	期望寿命	患病率	健康生存人年数	累积健康生存人年数	健康期望寿命	患病损失健康期望寿命
i	a_X	P_X	D_X	m_X	l_X	L_X	e_X	R_X	h_X	H_X	E_X	DLE_X
1	0.50	96211	4691	0.048757	100000	97620	70.77	0.053	92446	6097934	60.98	9.79
1	0.5	94971	674	0.007097	95240	94904	73.28	0.053	89874	6005487	63.06	10.23
1	0.5	94386	496	0.005255	94567	94319	72.80	0.053	89320	5915614	62.55	10.25
1	0.5	93936	405	0.004311	94071	93869	72.18	0.053	88894	5826294	61.93	10.25
1	0.5	93609	249	0.002660	93666	93542	71.49	0.053	88584	5737400	61.25	10.24
1	0.5	93379	212	0.002270	93418	93312	70.68	0.062	87526	5648816	60.47	10.21
1	0.5	93176	194	0.002082	93206	93109	69.84	0.067	86871	5561289	59.67	10.17
1	0.5	93015	127	0.001365	93012	92948	68.98	0.076	85884	5474419	58.86	10.13
1	0.5	92882	139	0.001497	92885	92816	68.08	0.077	85669	5388534	58.01	10.06
1	0.5	92766	95	0.001024	92746	92699	67.18	0.104	83058	5302866	57.18	10.00
1	0.5	92632	172	0.001857	92651	92565	66.25	0.123	81180	5219808	56.34	9.91
1	0.5	92499	95	0.001027	92479	92432	65.37	0.151	78475	5138628	55.57	9.80
1	0.5	92399	107	0.001158	92384	92331	64.44	0.174	76265	5060153	54.77	9.66
1	0.5	92301	88	0.000953	92277	92233	63.51	0.175	76093	4983888	54.01	9.50
1	0.5	92194	127	0.001378	92189	92126	62.57	0.178	75728	4907795	53.24	9.33
1	0.5	92070	121	0.001314	92063	92002	61.66	0.236	70290	4832068	52.49	9.17
1	0.5	91941	138	0.001501	91942	91873	60.74	0.293	64954	4761778	51.79	8.94
1	0.5	91788	166	0.001809	91804	91721	59.83	0.344	60169	4696824	51.16	8.67
1	0.5	91633	145	0.001582	91638	91565	58.93	0.425	52650	4636655	50.60	8.34
1	0.5	91479	163	0.001782	91493	91412	58.03	0.053	86567	4584005	50.10	7.92
1	0.5	91295	205	0.002245	91330	91228	57.13	0.053	86393	4497439	49.24	7.89

图 13-10　“已知人口与死亡数”工作表的健康期望寿命表

第 14 章

随机化分组与样本含量估计

为了控制和降低系统误差对实验结果的影响，合理估计抽样误差，更确切地回答研究者事先提出的假说，有必要进行实验设计。实验设计时需要遵循对照原则（rule of control）、随机化原则（rule of randomization）和重复原则（rule of replication）等基本原则。本章将针对后两个原则，分别介绍 Excel 用于随机化分组与样本含量估计的方法。

14.1 随机化分组

随机化是采用随机的方法使每个受试对象都有同等机会进入试验组与对照组，使非处理因素以相等的机会影响各组，以避免非处理因素带来的偏倚。随机化方法有：抛硬币、掷骰子、拈阄、随机数字表、电脑随机函数等，本节将介绍如何使用 Excel 进行随机化。

常用的实验设计方法有完全随机设计、配对设计、配伍组设计等。

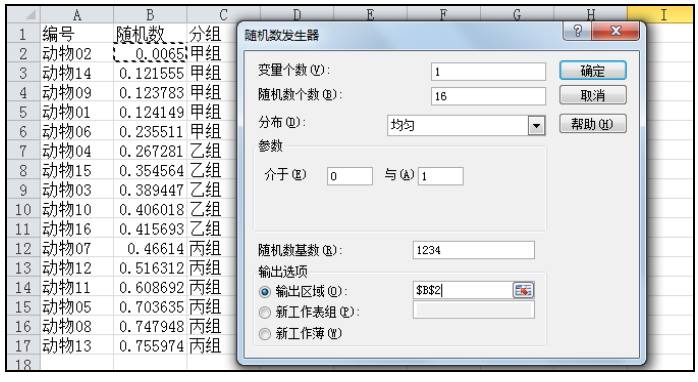
14.1.1 完全随机设计

完全随机设计（completely randomized design）是将受试对象编号后，采用随机的方法使受试对象具有同等机会被分配到各个处理组中。每个处理组的例数可以相等，也可以不相等，但各组例数相等可增加样本的利用效率。当观测的数据为定量数据时，两个处理组常用两个独立样本 t 检验，两个以上处理组常采用完全随机设计的方差分析来进行假设检验（见第 8 章）。

例 14-1：用电脑产生随机数，将 16 只动物随机分配到甲、乙、丙三组中，甲、乙两组各 5 只，丙组分 6 只。

解：按如下步骤完成随机化分组。

①将动物编号，分别编为动物 01、动物 02、……、动物 16（如图 14-1 所示，也可见光盘文件“14-1 随机化分组.xlsx”的“完全随机设计”工作表）。事先规定：随机数按从小到大排序后，前 5 个较小随机数相应编号的动物分配到甲组，最后 6 个较大随机数相应编号的动物分配到丙组，其余 5 个（中间）相应编号的动物分配到乙组。



	A	B	C
1	编号	随机数	分组
2	动物02	0.0065	甲组
3	动物14	0.121555	甲组
4	动物09	0.123783	甲组
5	动物01	0.124149	甲组
6	动物06	0.235511	甲组
7	动物04	0.267281	乙组
8	动物15	0.354564	乙组
9	动物03	0.389447	乙组
10	动物10	0.406018	乙组
11	动物16	0.415693	乙组
12	动物07	0.46614	丙组
13	动物12	0.516312	丙组
14	动物11	0.608692	丙组
15	动物05	0.703635	丙组
16	动物08	0.747948	丙组
17	动物13	0.755974	丙组
18			

图 14-1　完全随机设计的随机化分组

②在单元格 A1、B1、C1 中分别输入文字“编号”、“随机数”、“分组”；在单元格 A2

中编号“动物 01”，然后将单元格 A2 中的公式向下拖放填充至单元格 A17。

③为了使产生的随机数不变动，不采用随机函数，而选择功能区“数据→分析→数据分析→随机数发生器”命令，在“随机发生数”对话框的“变量个数”后填“1”，“随机数个数”后填“16”，分布选择“均匀”，“随机数基数”处任给一个数（如 1234，不同基数将产生不同的随机数），“输出区域”选择 B2。单击“确定”按钮。

④将鼠标置于 B1 单元格，单击功能区“数据→排序”命令，在“排序”对话框中的“主要关键字”选择“随机数”，“次序”选择“升序”，勾选“数据包含标题”，单击“确定”按钮（或选择单元格 B1 后，选择功能区“开始→排序和筛检→升序”命令）。

⑤按事先规定分配 16 只动物。在单元格 C2 中输入文字“甲组”，将鼠标移至单元格 C2 的右下角填充柄处，向下拖放填充至 C6；在单元格 C7 中输入文字“乙组”，将鼠标移至单元格 C7 的右下角填充柄处，向下拖放填充至 C11；在单元格 C12 中输入文字“丙组”，将鼠标移至单元格 12 的右下角填充柄处，向下拖放填充至 C17。随机化分组结果为：编号为动物 02、14、09、01、06 的动物被分配到甲组；编号为动物 04、15、03、10、16 的动物被分配到乙组；编号为动物 07、12、11、05、08、13 的动物被分配到丙组。

应特别注意上述的随机数基数（random seed）为 1234，不同的随机数基数会得到不同的随机分组。

14.1.2 配对设计

配对设计（matched-pairs design）是将受试者按一定条件配成对子，对子中的一个为“试验组”，另一个为“对照组”，分别接受不同的处理。每个处理组的例数相等。一般以非实验因素，如性别、年龄、动物的窝别等作为配对条件。配对资料的 t 检验（见第 8 章）、配对 χ^2 检验（见第 9 章）可用来进行统计学假设检验。

例 14-2：取 20 只小白鼠，按体重、窝别配成 10 对，采用 Excel 随机数发生器将其分配到甲、乙组中。

解：按如下步骤完成随机化分组。

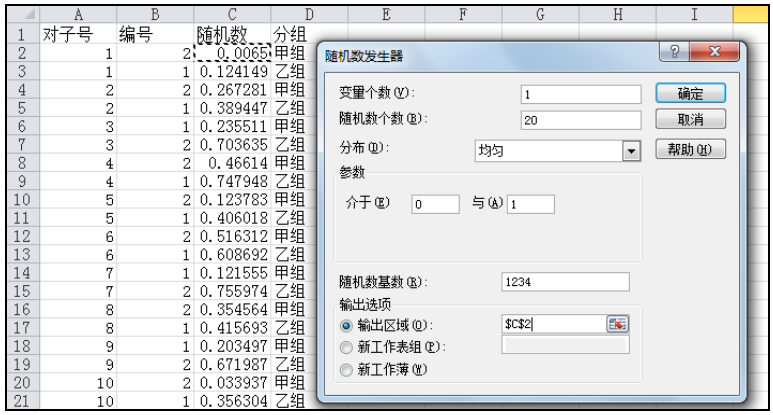
①将动物编号，并给出分配动物的规定。先将 10 对小白鼠编上小鼠 1、小鼠 2、……、小鼠 10 对子号，再给每对动物分别编上 1、2 号。规定：在每个对子中，随机数较小者为甲组，随机数较大者为乙组。

②在单元格 A1、B1、C1、D1 中分别输入文字“对子号”、“编号”、“随机数”、“分组”；在单元格 A2、A3、A4、A5、……、A21 中分别输入数字 1、1、2、2、……、10；在单元格 B2、B3、B4、B5、……、B21 中分别输入数字 1、2、1、2、……、2（如图 14-2 所示，也可见光盘文件“14-1 随机化分组.xlsx”的“配对设计”工作表）。

③选择功能区“数据→分析→数据分析→随机数发生器”命令，在“随机数发生器”对话框的“变量个数”后填“1”，在“随机数个数”后填“20”，分布选择“均匀”，“随机数基数”处任给一个数（本例为 1234，不同基数将产生不同的随机数），“输出区域”选择 C2。单击“确定”按钮。

④单击功能区“数据→排序”命令，“主要关键字”选择“对子号”，“次序”选择“升序”；“次要关键字”选择“随机数”，“次序”选择“升序”。勾选“数据包含标题”，单击“确定”按钮。

⑤按事先规定分配每个对子中的 1、2 号分别进入甲、乙组。在单元格 D2、D3、D4、D5、……、D21 中分别输入文字“甲组”、“乙组”、“甲组”、“乙组”、……“乙组”。分组结果及“随机数发生器”对话框如图 14-2 所示。



	A	B	C	D
1	对子号	编号	随机数	分组
2	1	2	0.00655	甲组
3	1	1	0.124149	乙组
4	2	2	0.267281	甲组
5	2	1	0.389447	乙组
6	3	1	0.235511	甲组
7	3	2	0.703635	乙组
8	4	2	0.46614	甲组
9	4	1	0.747948	乙组
10	5	2	0.123783	甲组
11	5	1	0.406018	乙组
12	6	2	0.516312	甲组
13	6	1	0.608692	乙组
14	7	1	0.121555	甲组
15	7	2	0.755974	乙组
16	8	2	0.354564	甲组
17	8	1	0.415693	乙组
18	9	1	0.203497	甲组
19	9	2	0.671987	乙组
20	10	2	0.033937	甲组
21	10	1	0.356304	乙组

图 14-2　配对设计的随机化分组

分组结果是：对子号 1、2、3、4、5、6、7、8、9、10 中相应编号 2、2、1、2、2、2、1、2、1、2 的动物为甲组；相应编号 1、1、2、1、1、1、2、1、2、1 的动物为乙组。

14.1.3　配伍组设计

配伍组设计又称随机区组设计（randomized and block design），实际上是配对设计的扩展，此时处理组不再只是两组而是两组以上，配对设计的对子号相当于配伍组区组编号。将受试者按一定条件配成配伍组（或区组），每个配伍组中有 3 个或 3 个以上受试对象，分别随机分配到各个处理组中。每个处理组的例数相等。一般以非实验因素，如性别、年龄、动物的窝别等作为配伍组条件。随机区组设计的方差分析（见第 8 章的 8.7 节）可用来进行统计学假设检验。

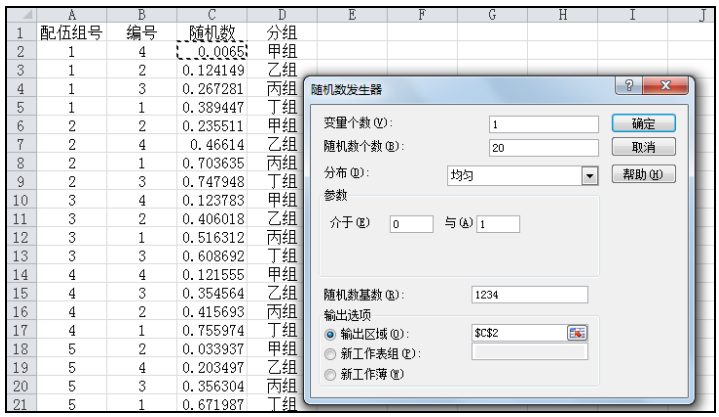
例 14-3：将体重相近的 20 只小鼠配成 5 个区组，每个区组内的 4 只小鼠分别随机

分配到甲、乙、丙、丁四组中，喂以四种不同饲料。采用 Excel 随机数发生器将每个区组内的小鼠进行随机分配。

解：配伍组设计的随机化分组与配对设计十分相似，具体操作步骤如下：

①将动物编号，并给出分配动物的规定。先将 5 个配伍组小鼠编上小鼠 1、2、……、5 号，再给每个配伍组中 4 只小鼠分别编上 1、2、3、4 号。规定：在每个配伍组中，随机数从小到大排列依次分配到甲组、乙组、丙组、丁组中。

②在单元格 A1、B1、C1、D1 中分别输入文字“配伍组号”、“编号”、“随机数”、“分组”；在单元格 A2、A3、A4、A5、……、A21 中分别输入数字 1、1、1、1、……、5；在单元格 B2、B3、B4、B5、……、B21 中分别输入数字 1、2、3、4、……、4（如图 14-3 所示，也可见光盘文件“14-1 随机化分组.xlsx”的“配伍组设计”工作表）。



	A	B	C	D
1	配伍组号	编号	随机数	分组
2	1	4	0.0065	甲组
3	1	2	0.124149	乙组
4	1	3	0.267281	丙组
5	1	1	0.389447	丁组
6	2	2	0.235511	甲组
7	2	4	0.46614	乙组
8	2	1	0.703635	丙组
9	2	3	0.747948	丁组
10	3	4	0.123783	甲组
11	3	2	0.406018	乙组
12	3	1	0.516312	丙组
13	3	3	0.608692	丁组
14	4	4	0.121555	甲组
15	4	3	0.354564	乙组
16	4	2	0.415693	丙组
17	4	1	0.755974	丁组
18	5	2	0.033937	甲组
19	5	4	0.203497	乙组
20	5	3	0.356304	丙组
21	5	1	0.671987	丁组

图 14-3　配伍组设计的随机化分组

③单击功能区“数据→分析→数据分析→随机数发生器”，在“随机数发生器”对话框的“变量个数”后填“1”，“随机数个数”后填“20”，分布选择“均匀”，“随机数基数”处任给一个数（本例为 1234，不同基数将产生不同的随机数），“输出区域”选择 C2。单击“确定”按钮。

④选取单元格区域 A1:C21，单击功能区“数据→排序”命令，“主要关键字”选“配伍组号”，“次序”选择“升序”；“次要关键字”选择“随机数”，“次序”选择“升序”，勾选“数据包含标题”，单击“确定”按钮。

⑤按事先规定分配每个配伍组中的 1、2、3、4 号分别进入甲、乙、丙、丁组。在单元格区域 D2:D5、D6:D9、D10:D13、D14:D17、D18:D21 中分别输入文字“甲组”、“乙组”、“丙组”、“丁组”。分组结果及随机数发生器对话框如图 14-3 所示。

分组结果是：配伍组号 1、2、3、4、5 中相应的编号 4、2、4、4、2 被分配到甲组；相应编号 2、4、2、3、4 被分配到乙组；相应编号 3、1、1、2、3 被分配到丙组；相应编

号 1、3、3、1、1 被分配到丁组。

14.1.4 三个随机化分组程序介绍

光盘文件“14-2 完全随机分组.xlsx”、“14-3 配对设计分组.xlsx”、“14-4 随机区组分组.xlsx”分别可简单解决完全随机设计样本分组、配对设计样本分组、随机区组设计样本分组问题。

对于例 14-1，可打开光盘文件“14-2 完全随机分组.xlsx”，在单元格 B4 中键入组数为“3”；在单元格 B9 中输入“动物 1”，将此单元格名称向下拖放至 B24，此时会在单元格区域 F9:G208 范围内显示分组结果（如图 14-4 所示）。如将动物 12、16、13、3、9 分配到 1 组，动物 5、15、1、10、2 分配到 2 组，动物 7、8、6、4、11、14 分配到 3 组。因为程序采用“rand()”函数编制，不能固定给出随机数基数，所以每次运行都会产生不同的分组方案。该程序最大可对 200 个个体进行随机化分组。如某研究生班共计 198 人，拟随机分成 10 个组进行实验课考试，用户只需将 198 个研究生的姓名输入到该程序的 B9:B208 范围内，改变组数为 10 即可。

	A	B	C	D	E	F	G
1	完全随机分组						
2							
3		组数				总例数	16
4		3				每组最多个体数	5
5						每组最少个体数	6
6							
7						分组结果	
8	编号	输入个体名称	随机数	排序		个体名称	组别
9	1	动物1	0.739224	5		动物5	1
10	2	动物2	0.564933	8		动物8	1
11	3	动物3	0.56941	7		动物7	1
12	4	动物4	0.320486	14		动物14	1
13	5	动物5	0.138473	15		动物15	1
14	6	动物6	0.876525	1		动物1	2
15	7	动物7	0.861757	2		动物2	2
16	8	动物8	0.459954	12		动物12	2
17	9	动物9	0.762215	4		动物4	2
18	10	动物10	0.520363	11		动物11	2
19	11	动物11	0.069935	16		动物16	3
20	12	动物12	0.537539	9		动物9	3
21	13	动物13	0.724695	6		动物6	3
22	14	动物14	0.530271	10		动物10	3
23	15	动物15	0.77669	3		动物3	3
24	16	动物16	0.422732	13		动物13	3

图 14-4 完全随机设计分组程序界面

对于例 14-2，可打开光盘文件“14-3 配对设计分组.xlsx”，在单元格 B7 中键入“老鼠 1”，将此单元格名称向右拖放至 C7（变成“老鼠 2”），然后选取 B7:C7 向下拖放至 C16，此时会在单元格区域 M5:O16 范围内显示分组结果（如图 14-5 所示）。因为程序采用“rand()”函数编制，不能固定给出随机数基数，所以每次运行都会产生不同的分组方案。该程序最大可对 100 个对子进行随机化分组。

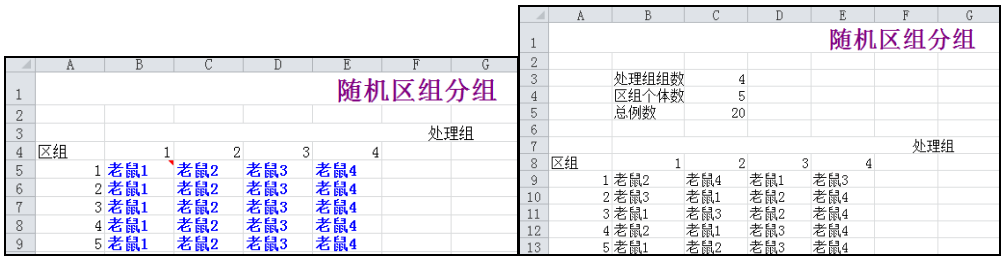


配对设计分组

												对子数	10	
	每一对子内编号			每一对子内编号			每一对子内编号					分组结果		
	对子	1	2	对子	1	2	对子	1	2		对子	A	B	
7	1	老鼠1	老鼠2	1	0.256909	0.907176	1	2	1		1	老鼠2	老鼠1	
8	2	老鼠1	老鼠2	2	0.68743	0.797282	2	2	1		2	老鼠2	老鼠1	
9	3	老鼠1	老鼠2	3	0.105673	0.826409	3	2	1		3	老鼠2	老鼠1	
10	4	老鼠1	老鼠2	4	0.588342	0.856207	4	2	1		4	老鼠2	老鼠1	
11	5	老鼠1	老鼠2	5	0.396615	0.062234	5	1	2		5	老鼠1	老鼠2	
12	6	老鼠1	老鼠2	6	0.303969	0.885042	6	2	1		6	老鼠2	老鼠1	
13	7	老鼠1	老鼠2	7	0.485279	0.288406	7	1	2		7	老鼠1	老鼠2	
14	8	老鼠1	老鼠2	8	0.766942	0.775988	8	2	1		8	老鼠2	老鼠1	
15	9	老鼠1	老鼠2	9	0.485613	0.613166	9	2	1		9	老鼠2	老鼠1	
16	10	老鼠1	老鼠2	10	0.358492	0.757392	10	2	1		10	老鼠2	老鼠1	

图 14-5 配对设计分组程序界面

对于例 14-3，可打开光盘文件“14-4 随机区组分组.xlsx”的“数据与计算”工作表（如图 14-6 左侧所示），在单元格 B5 中键入“老鼠 1”，将此单元格名称向右拖放至 E5（变成“老鼠 4”），然后选取 B5:E5 向下拖放至 E9，此时会在“结果”工作表中显示分组结果（如图 14-6 右侧所示）。因为程序采用“rand()”函数编制，不能固定给出随机数基数，所以每次运行都会产生不同的分组方案。该程序最大可对 100 个区组，每区组内 10 个个体进行随机化分组。



随机区组分组

				处理组	
区组	1	2	3	4	
1	老鼠1	老鼠2	老鼠3	老鼠4	
2	老鼠1	老鼠2	老鼠3	老鼠4	
3	老鼠1	老鼠2	老鼠3	老鼠4	
4	老鼠1	老鼠2	老鼠3	老鼠4	
5	老鼠1	老鼠2	老鼠3	老鼠4	

随机区组分组

处理组组数	4				
区组个体数	5				
总例数	20				
					处理组
区组	1	2	3	4	
1	老鼠2	老鼠4	老鼠1	老鼠3	
2	老鼠3	老鼠1	老鼠2	老鼠4	
3	老鼠1	老鼠3	老鼠2	老鼠4	
4	老鼠2	老鼠1	老鼠3	老鼠4	
5	老鼠1	老鼠2	老鼠3	老鼠4	

图 14-6 随机区组设计的分组程序界面



注意：程序的优点是获得结果比较快，但由于采用的“rand()”没有随机数基数，其他人无法重复验证你的分组结果。

14.2 样本含量的估计

重复是指各实验处理组（或观察研究组）要有一定的重复数量，即样本含量（sample size）。样本含量越大，抽样误差越小（如反映抽样误差大小的标准误有：标准误 = 标准差/$\sqrt{样本含量}$），研究结果的精确度（precision）越高，从而可提高研究的准

确度（accuracy，样本估计值与总体真实值之间的接近程度）。样本含量过多（甚至包括研究整个总体），将加大实验规模，延长实验时间，浪费人力、财力、物力，甚至有可能间接增加系统误差；样本含量过少，抽样误差就会很大，最终结果将无法达到研究的目的。因此，有必要对研究所需最少样本含量进行估计。

样本含量一般由以下 4 个因素决定。

① 检验水准（significance level，α）。这是拒绝正确的零假设 H_0 的概率，也称 I 型错误的概率，是研究前人为事先设定的具有统计学意义的最大 P 值，通常取 0.05 或 0.01，可采用单侧或双侧求临界值。α越小，也就是具有统计学意义的 P 值设定越低，犯 I 类错误的概率越小，则所需的样本含量越大；双侧检验比单侧检验所需的样本含量要大。

② 检验功效（test power，$1-\beta$）。这是接受正确备择假设 H_1 的概率，其中β也称 II 型错误的概率，即当差异存在时检出这种差异的概率。通常要求大于 80%，常规取 90%或 95%，一般只采用单侧求临界值。应该牢记：如果检验功效较低，即使两组之间有差异存在，也不能够通过检验观察出这种差异。所以当检验功效较低时，就不值得再去做试验了。检验功效越大，所需样本含量越大。

③ 总体变异度（variability）。如与均数对应的标准差σ，以及与率对应的标准差 $\sqrt{\pi(1-\pi)}$。σ越大，即资料的离散程度或个体变异程度越大，所需的样本含量越大。定量数据个体值偏离均数越远或定性数据率偏离 0.5 越远（如图 14-7 所示），所需的样本含量越大。

④容许误差或有意义差值（significance difference）δ。如$\delta=\mu_1-\mu_2$或$\delta=\pi_1-\pi_2$。该差值是研究者认为重要、且不情愿忽略的值，临床医生通常认为舒张期差值=0.69kPa（5mmHg），白细胞差值=0.5×10^9/l（500 个/mm^3）是最小有意义差值。注意：不要把该差值与研究者期望从试验中得到的差值混为一团。容许误差越小，所需样本含量越大。

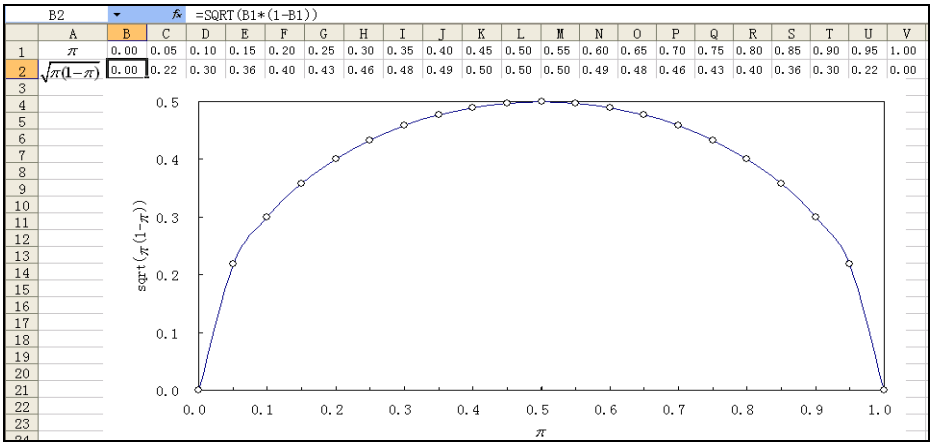


图 14-7　率与率的标准差之间的关系

总体参数通常无法知晓，人们往往通过查阅资料、借鉴前人经验或进行预实验来获取总体参数的估计值。

14.2.1 Nomogram 图示法

Altman，DG（British Medical Journal, 281, 1336-1338）于 1980 年提出了采用图示法（如图 14-8 所示）获得所需最低样本含量，该方法主要适用于配对 t 检验、两组 t 检验和两组率的比较。图中的标准化差值是容许误差 δ 与总体变异度 σ 的比值。配对 t 检验、两组 t 检验和两组率的比较的标准化差值分别为：

$$\frac{2\delta}{\sigma_d}=\frac{2\text{最小有意义对子间差值}}{\text{对子间差值的标准差}}$$

$$\frac{\delta}{\sqrt{\sigma^2_{\mu_1-\mu_2}}}=\frac{\text{两组间有意义差值}}{\text{两组的合并标准差}}$$

$$\frac{\pi_1-\pi_2}{\sigma_{\pi_1-\pi_2}}\approx\frac{p_1-p_2}{\sqrt{\frac{p_1+p_2}{2}\left(1-\frac{p_1+p_2}{2}\right)}}$$

获得的样本含量分别是对子数、两组总例数。

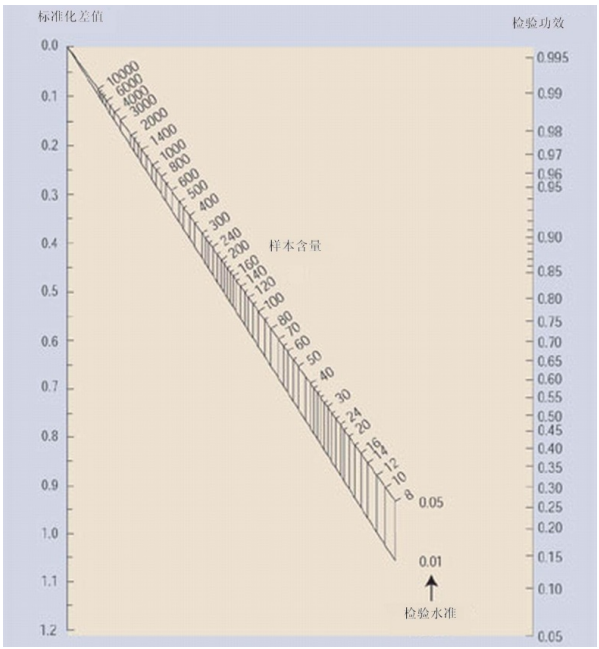


图 14-8 Nomogram 图示法获取样本含量

例如，对于配对资料，$\alpha = 0.05$，检验功效=90%，差值的标准差= 4.6，$\delta = 1.5$。由此计算获得标准化差值= 2 (1.5) / (4.6) =0.65，利用图 14-8 查得样本含量为 100，即总共 100 个对子。

对于成组资料，α=0.05，检验功效=85%，两组标准差 s=0.44（从预试验中得到），δ=0.4。由此计算获得标准化差值= 0.4 / 0.44 = 0.91，利用图 14-8 查得样本含量为 45，即每组约需要的最低样本含量为 23。

对于两个率比较资料，$\alpha = 0.05$，检验功效=80%，$p_1 = 40\%$，$p_2 = 30\%$。由此计算获得：

$$\text{标准化差值}=\frac{0.4-0.3}{\sqrt{\frac{0.4+0.3}{2}\left(1-\frac{0.4+0.3}{2}\right)}}=\frac{0.1}{\sqrt{0.35\times 0.65}}=0.21$$

利用图 14-8 查得样本含量为 700，即每组各需 350 个个体。

14.2.2 假设检验的样本含量估计

不同的研究设计、不同的数据类型、不同的统计学方法将需要不同的样本含量。下面将按假设检验的样本含量估计、置信区间的样本含量估计、特殊设计的样本含量估计等进行分别叙述。

光盘文件“14-5 假设检验样本含量.xlsx”中包含 6 个工作表，可对单个或成组均数比较、单个或成组率比较以及简单相关分析进行样本含量估计。

1. 定量数据单组与配对设计

例 14-4：已知血吸虫病人血红蛋白平均含量为 9g/100ml，标准差为 2.5g/100ml，如果采用呋喃丙胺治疗后能使血红蛋白平均含量增加至 10g/100ml，则认为治疗有效。问在检验水准 α=0.05，检验功效 1–β=90%情况下，该研究至少需要治疗多少病人？

解：根据题意，血红蛋白平均含量是定量数据，且是比较接受治疗的血吸虫病患者血红蛋白均数 μ 是否高于一般未接受治疗血吸虫病患者血红蛋白均数 μ_0，即 H_0：$\mu = \mu_0$，H_1：$\mu > \mu_0$，所以检验取单侧。

本例已知检验水准 $\alpha = 0.05$，检验功效 $1-\beta = 0.90 = 90\%$，变异指标（即总体标准差 σ 的估计值）$S = 2.5$ g/100ml，容许误差 $\delta = \mu - \mu_0 = 10 - 9 = 1$（g/100ml），将此代入公式：

$$n=\left[\frac{(t_{\alpha/2,\nu}+t_{\beta,\nu})S}{\delta}\right]^2 \qquad (14\text{-}1)$$

便可获得所需的双侧最低样本含量。式中，$t_{\alpha/2,\nu}$、$t_{\beta,\nu}$ 与 $\nu = n-1$ 有关，分别为 I 型错误概率 α 相应的双侧 t 界值、II 型错误概率 β 相应的单侧 t 界值。在实际计算时，先用自由度 $\nu = \infty$ 时（即 $t_{\alpha/2,\infty} = z_{\alpha/2}$、$t_{\beta/2,\infty} = z_{\beta/2}$）粗略估计样本含量 n^1；第二次用 $\nu = n^1 - 1$ 时

的$t_{\alpha/2,\nu}$和$t_{\beta,\nu}$，代入公式（14-1）中粗略估计样本含量n^2；第三次用$\nu=n^2-1$时的$t_{\alpha/2,\nu}$和$t_{\beta,\nu}$，代入公式（14-1）中粗略估计样本含量n^3；……；直至相邻两个样本含量结果趋于稳定为止。

根据以上原理编制的样本含量计算模板如图 14-9 所示。左上侧为数据输入区，右上侧为结果显示区，其下方是有关计算公式与中间结果区。

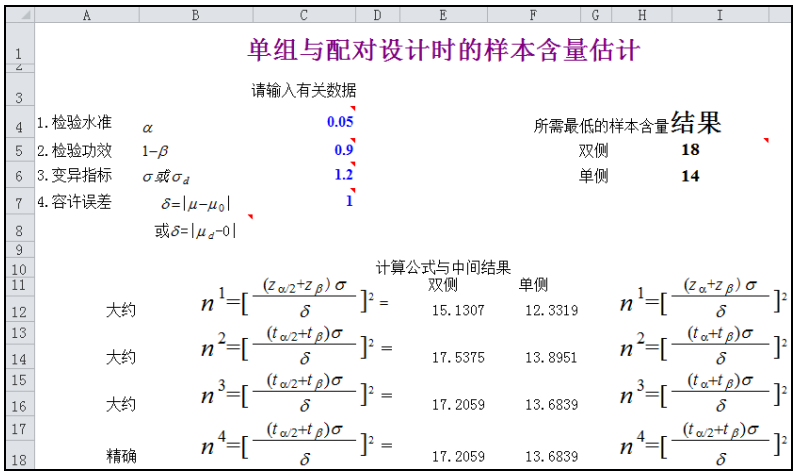


图 14-9 样本均数与总体均数比较的样本含量估计

结果分析：将已知数据输入程序界面后，获得单侧所需最低样本含量为 55，即该研究至少需要治疗 55 名血吸虫病病人

温馨提示：用户只需修改图 14-9 中单元格区域 C4:C7 中的数据，便可获得样本均数与总体均数比较所需的样本含量。

配对定量资料比较的样本含量估计公式和上面介绍的样本均数与总体均数比较相同。不同的是：配对设计的变异指标为对子差值的标准差，获得的样本含量为对子数。

例 14-5：如果治疗后比治疗前白细胞平均改变 1×10^9/l，则可认为碳酸铝对白细胞减少症的治疗有效。由以往资料得知，相应的白细胞标准差为 1.2×10^9/l。试估计在α=0.05、1−β=90%情况下，至少应治疗多少病人？

解：配对资料比较的计算方法与“样式均数与总体均数比较”相同，其H：$\mu-\mu=\mu=0$，H_1：$\mu_d\neq0$。将检验水准$\alpha=0.05$、检验功效$1-\beta=0.90=90\%$、变异指标（即总体差值标准差的估计值）$\hat{\sigma}_d=S_d=1.2$（10^9/l）、容许误差$\delta=\mu_d-0=1-0=1$（10^9/l）（d为两组差值）替换单元格区域 C4:C7 中的数据，获得双侧所需最低样本含量为 18，单侧所

需最低样本含量为 15。即该研究至少需要治疗 18 对白细胞减少症病人。

2. 定量数据成组设计

两样本均数比较时所需双侧样本含量的计算公式为：

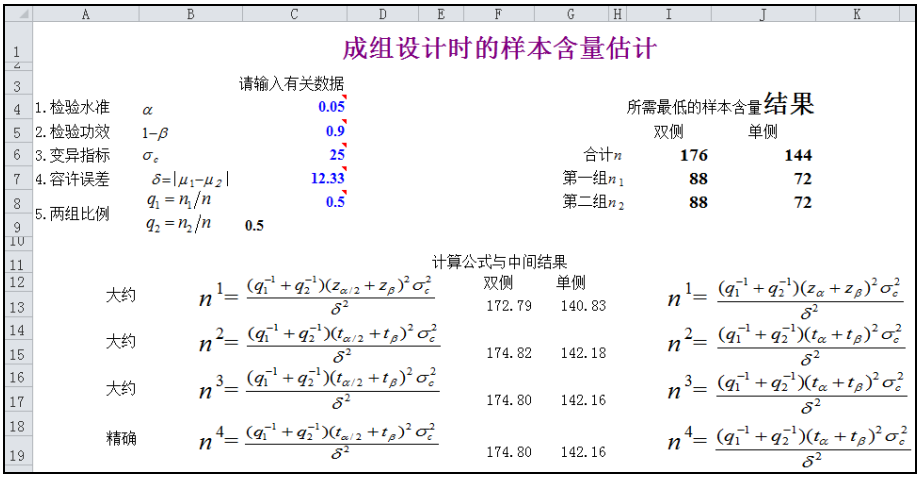
$$n=\frac{\left(q_1^{-1}+q_2^{-1}\right)(t_{\alpha/2,\nu}+t_{\beta,\nu})^2 S_C^2}{\delta^2} \qquad (14\text{-}2)$$

公式中，$n=n_1+n_2$ 为两组共需的样本含量，n_1、n_2 分别为第一组、第二组所需的样本含量，两组样本所占比例分别为 $q_1=n_1/n$、$q_2=n_2/n=1-q_1$。S_C 为两个总体合并标准差的估计值（参见 8.4.2 节）；δ 为两样本对应总体均数之差，即 $\delta=|\mu_1-\mu_2|$，此差值是专业上认为有重要意义的最小差值；$t_{\alpha/2,\nu}$、$t_{\beta,\nu}$ 与 $\nu=n_1+n_2-2=n-2$ 有关，分别为 I 型错误概率 α 与 II 型错误概率 β 相应的 t 界值。公式（14-2）与公式（14-1）十分相似，也需要先用自由度 $\nu=\infty$ 时（即 $t_{\alpha/2,\infty}=z_{\alpha/2}$、$t_{\beta/2,\infty}=z_{\beta/2}$）粗略估计样本含量 n^1；第二次用 $\nu=n^1-2$ 时的 $t_{\alpha/2,\nu}$ 和 $t_{\beta,\nu}$，代入公式（14-2）中粗略估计样本含量 n^2；第三次用 $\nu=n^2-2$ 时的 $t_{\alpha/2,\nu}$ 和 $t_{\beta,\nu}$，代入公式（14-2）中粗略估计样本含量 n^3；……；直至相邻两个样本含量结果趋于稳定为止。

例 14-6：在动物镇咳实验中，由以往文献得知中药复方 I 平均推迟小白鼠咳嗽发生时间为 33.67 秒，复方 II 平均推迟小白鼠咳嗽发生时间为 46.00 秒。设两组标准差均为 25 秒，α=0.05，1−β=0.90，问至少需要用多少只小白鼠做实验可获得有意义结果？

解：此数据为两样本均数比较，其 H_0：$\mu_1=\mu_2$，H_1：$\mu_1\neq\mu_2$。打开光盘文件“14-5 假设检验样本含量.xlsx”的“定量数据成组设计”工作表，将检验水准 $\alpha=0.05$、检验功效 $1-\beta=0.90=90\%$、变异指标（即总体合并标准差的估计值）$\hat{\sigma}_C=S_C=25$（秒）（当两组例数、标准差均相同时，则合并标准差等于各组的标准差）、容许误差 $\delta=|\mu_1-\mu_2|$=46-33.67=12.33（秒）、$q_1=q_2=0.5$（没有给出两组样本含量比例情况下，采用两组样本含量相等）替换如图 14-10 所示单元格区域 C4:C8 中的数据，获得各组双侧所需的最低样本含量为 88，单侧所需的最低样本含量为 72。即该研究至少需要 176 只老鼠做试验，才能得出有意义的结果。如果进行单侧检验，则只需 144 只老鼠（如图 14-10 所示）。

温馨提示：用户只需修改如图 14-10 所示的单元格区域 C4:C8 中的数据，便可获得其他两样本均数比较所需的最低样本含量。



成组设计时的样本含量估计

请输入有关数据

			所需最低的样本含量结果	双侧	单侧
1. 检验水准	α	0.05			
2. 检验功效	$1-\beta$	0.9			
3. 变异指标	σ_c	25	合计n	176	144
4. 容许误差	$\delta=\lvert\mu_1-\mu_2\rvert$	12.33	第一组n_1	88	72
5. 两组比例	$q_1=n_1/n$	0.5	第二组n_2	88	72
	$q_2=n_2/n$	0.5			

计算公式与中间结果

		双侧	单侧	
大约	$n^1=\frac{(q_1^{-1}+q_2^{-1})(z_{\alpha/2}+z_\beta)^2\sigma_c^2}{\delta^2}$	172.79	140.83	$n^1=\frac{(q_1^{-1}+q_2^{-1})(z_\alpha+z_\beta)^2\sigma_c^2}{\delta^2}$
大约	$n^2=\frac{(q_1^{-1}+q_2^{-1})(t_{\alpha/2}+t_\beta)^2\sigma_c^2}{\delta^2}$	174.82	142.18	$n^2=\frac{(q_1^{-1}+q_2^{-1})(t_\alpha+t_\beta)^2\sigma_c^2}{\delta^2}$
大约	$n^3=\frac{(q_1^{-1}+q_2^{-1})(t_{\alpha/2}+t_\beta)^2\sigma_c^2}{\delta^2}$	174.80	142.16	$n^3=\frac{(q_1^{-1}+q_2^{-1})(t_\alpha+t_\beta)^2\sigma_c^2}{\delta^2}$
精确	$n^4=\frac{(q_1^{-1}+q_2^{-1})(t_{\alpha/2}+t_\beta)^2\sigma_c^2}{\delta^2}$	174.80	142.16	$n^4=\frac{(q_1^{-1}+q_2^{-1})(t_\alpha+t_\beta)^2\sigma_c^2}{\delta^2}$

图 14-10 两样本均数比较时样本含量的估计

3. 定性数据单组设计

当总体率小于 0.2 或大于 0.8（极度偏离 0.5）时，样本率与总体率比较所需双侧样本含量的计算公式为：

$$n=0.25\left[\frac{z_{\alpha/2}+z_\beta}{\sin^{-1}\sqrt{p}-\sin^{-1}\sqrt{\pi}}\right]^2 \qquad (14\text{-}3)$$

公式中，n为所需的样本含量，p为样本率，π为总体率，sin (•) 为反正弦函数，$z_{\alpha/2}$和z_β分别为 I 型错误概率α与 II 型错误概率β相应的标准正态离差界值。

当总体率$0.2\leqslant\pi\leqslant0.8$时，样本率与总体率比较的双侧样本含量计算公式为：

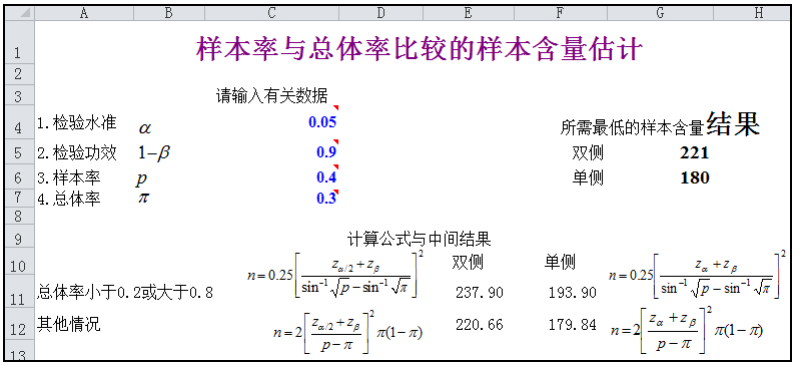
$$n=2\left[\frac{z_{\alpha/2}+z_\beta}{p-\pi}\right]^2\pi(1-\pi) \qquad (14\text{-}4)$$

式中符号含义与公式（14-3）相同。

例 14-7：一般药治疗慢性肾炎的有效率只有 30%，如果新药有效率为 40%方可认为有临床意义。问在α=0.05、1−β=0.90 条件下，需要多少研究病人？

解：该例为样本率p与总体率π的比较问题，H_0：$\pi_1=\pi$，H_1：$\pi_1\neq\pi$（这里π_1为样本率p所对应的总体率）。打开光盘文件“14-5 假设检验样本含量.xlsx”的“定性数据单组设计”工作表，将检验水准$\alpha=0.05$、检验功效$1-\beta=0.90=90\%$、样本率$p=0.4$、总体率$\pi=0.3$替换如图 14-11 所示的中单元格区域 C4:C7 中的数据，获得双侧所需的最低样本含量为 221，单侧所需的最低样本含量为 180。该例总体率$\pi=0.3$满足$0.2\leqslant\pi\leqslant0.8$，所以采用了公式（14-4）的计算结果。

温馨提示：用户只需修改如图 14-11 所示的单元格区域 C4:C7 中的数据，便可获得样本比较所需的最低样本含量。



样本率与总体率比较的样本含量估计

		请输入有关数据				
1. 检验水准	α	0.05		所需最低的样本含量结果		
2. 检验功效	$1-\beta$	0.9		双侧	221	
3. 样本率	p	0.4		单侧	180	
4. 总体率	π	0.3				
		计算公式与中间结果				
			双侧	单侧		
总体率小于0.2或大于0.8		$n=0.25\left[\frac{z_{\alpha/2}+z_{\beta}}{\sin^{-1}\sqrt{p}-\sin^{-1}\sqrt{\pi}}\right]^2$	237.90	193.90	$n=0.25\left[\frac{z_{\alpha}+z_{\beta}}{\sin^{-1}\sqrt{p}-\sin^{-1}\sqrt{\pi}}\right]^2$	
其他情况		$n=2\left[\frac{z_{\alpha/2}+z_{\beta}}{p-\pi}\right]^2\pi(1-\pi)$	220.66	179.84	$n=2\left[\frac{z_{\alpha}+z_{\beta}}{p-\pi}\right]^2\pi(1-\pi)$	

图 14-11　样本率与总体率比较时估计所需的样本含量

4. 定性数据配对设计

配对定性资料比较的资料格式见表 14-1。

表 14-1　配对定性资料比较的资料格式

甲　法	乙　法		合　计
	+	−	
+	a	b	$a+b$
−	c	d	$c+d$
合计	$a+c$	$b+d$	$n=a+b+c+d$

这种资料的双侧样本含量计算公式为：

$$n=\left[\frac{z_{\alpha/2}\sqrt{2\pi_c}+z_{\beta}\sqrt{\frac{2\pi_{-+}\pi_{+-}}{\pi_c}}}{\pi_{-+}-\pi_{+-}}\right]^2 \tag{14-5}$$

上述公式中，n 为所需的对子数；$\pi_{+-}=b/(a+b)$；$\pi_{-+}=c/(a+c)$；$\pi=(\pi_{+-}+\pi_{-+})/2$；$z_{\alpha/2}$ 和 z_{β} 分别为 I 型错误概率 α 与 II 型错误概率 β 相应的标准正态离差值。

例 14-8：已知甲、乙两种培养基接种细菌的 π_{+-}=0.04，$\pi_{-+}=0.25$。设定 α=0.05（双侧检验），$1-\beta$=0.90，问应采用多少样本对子数？

解：该例为配对设计四格表资料。打开光盘文件“14-5 假设检验样本含量.xlsx”的“定性数据配对设计”工作表，将检验水准 $\alpha=0.05$、检验功效 $1-\beta=0.90$、π_{+-}=0.04、

$\pi_{-+}=0.25$ 替换如图 14-12 所示的单元格区域 C4:C7 中的数据，获得双侧所需的最低样本含量为 54，单侧所需的最低样本含量为 43。

温馨提示：用户只需修改如图 14-12 所示的单元格区域 C4:C7 中的数据，便可获得定性资料配对设计所需的最低样本含量。在单元格区域 A12:E12 中输入配对四格表数据，可以获得 π_{+-}、π_{-+} 的估计值。

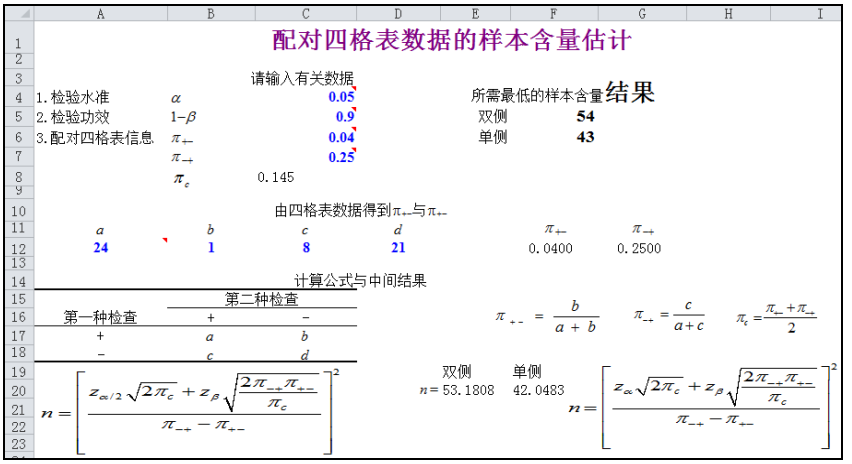


图 14-12 配对定性资料比较所需的样本含量估计

5. 定性数据成组设计

当两样本率比较时，双侧样本含量计算公式为：

$$n_1=\frac{[z_{\alpha/2}\sqrt{p(1-p)(1+c)/c}+z_{\beta}\sqrt{p_1(1-p_1)+p_2(1-p_2)/c}]^2}{(p_1-p_2)^2} \qquad (14\text{-}6)$$

式中，n_1 为样本 1 所需的样本含量，$c=n_2/n_1$，p_1、p_2 分别为样本 1、2 的样本率，$p=(p_1+cp_2)/(1+c)$，$z_{\alpha/2}$ 和 z_{β} 分别为 I 型错误概率 α 与 II 型错误概率 β 相应的标准正态离差值。

例 14-9：采用维生素 C 预防感冒，安慰剂组感冒发病率为 20%，预调查获得维生素 C 组的感冒发病率降为 10%。问在 α=0.05，1-β=0.90 条件下，研究应采用多大样本含量可获得维生素 C 组的感冒发病率具有统计学意义？

解：本例属于两样本率比较。H_0：$\pi_1=\pi_2$，H_1：$\pi_1\neq\pi_2$（双侧）或 $\pi_1>\pi_2$ 或 $\pi_1<\pi_2$（单侧）。打开光盘文件“14-5 假设检验样本含量.xlsx”的“定性数据成组设计”工作表，将检验水准 $\alpha=0.05$、检验功效 $1-\beta=0.90=90\%$、样本 1 的总体率估计值 p_1=0.2、样本

2 的总体率估计值 p_2=0.1、样本比例 c=1替换如图 14-13 所示的单元格区域 C4:C8 中的数据，获得第一、二组双侧所需的最低样本含量均为 266，单侧所需的最低样本含量均为 217。两组双侧、单侧共需最低样本含量分别为 532、434。

温馨提示：用户只需修改如图 14-13 所示的单元格区域 C4:C8 中的数据，便可获得两样本率比较所需的最低样本含量。

	A	B	C	D	E	F	G
1	定性数据成组设计时的样本含量估计						
2							
3			请输入有关数据			所需最低的样本含量结果	
4	1. 检验水准	α	0.05			双侧	单侧
5	2. 检验功效	$1-\beta$	0.9		合计	532	434
6	3. 样本率	p_1	0.2		第一组	266	217
7		p_2	0.1		第二组	266	217
8	4. 样本量比值	$c = n_2 / n_1$	1				
9	5. 合并率	p	0.15				
10							
11			计算公式与中间结果				
12		双侧检验：$n_1=\frac{[z_{\alpha/2}\sqrt{p(1-p)(1+c)/c}+z_{\beta}\sqrt{p_1(1-p_1)+p_2(1-p_2)/c}]^2}{(p_1-p_2)^2}$		=	265.855986		
13	第一组						
14		单侧检验：$n_1=\frac{[z_{\alpha}\sqrt{p(1-p)(1+c)/c}+z_{\beta}\sqrt{p_1(1-p_1)+p_2(1-p_2)/c}]^2}{(p_1-p_2)^2}$		=	216.49772		
15							

图 14-13　两样本率比较所需的样本含量估计

6. 简单线性相关分析

当进行简单线性相关分析时，双侧样本含量计算公式为：

$$n=4\left[\frac{z_{\alpha/2}+z_{\beta}}{\ln\left(\frac{1+\rho}{1-\rho}\right)}\right]^2+3 \qquad (14\text{-}7)$$

式中，ρ 为总体相关系数，$z_{\alpha/2}$ 和 z_{β} 分别为 I 型错误概率 α 与 II 型错误概率 β 相应的标准正态离差值。

例 14-10：估计总体相关系数估计值 $r=0.80$，问在单侧 α=0.05，1-β=0.90 条件下，研究应采用多大样本含量？

解：H_0：$\rho=0$，H_1：$\rho\neq0$。打开光盘文件“14-5 假设检验样本含量.xlsx”的“简单线性相关分析”工作表，将已知数据输入到单元格 C4、C5、C6（如图 14-14 所示）中，得到双侧样本含量为 12。即需最少观察 12 个个体，才能获得 $r=0.80$ 有统计学意义。

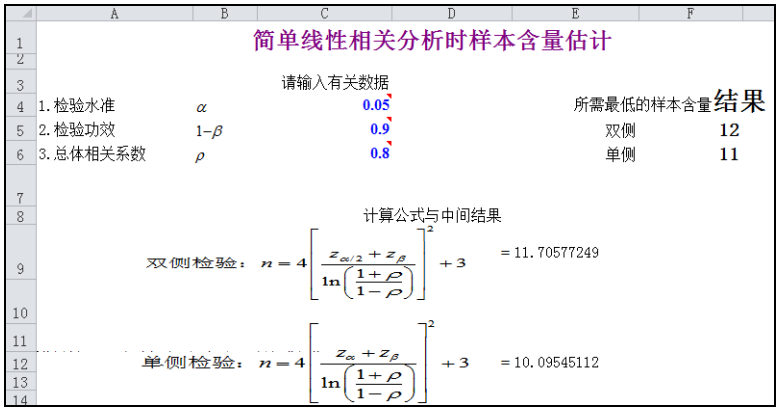


图 14-14 简单线性相关系数的样本含量估计

温馨提示：用户只需修改如图 14-14 所示的单元格区域 C4:C6 中的数据，便可获得简单线性相关分析所需的最低样本含量。

14.2.3 置信区间的样本含量估计

1. 总体均数的置信区间

如果n为样本含量，$\overline{X}$为总体均数μ的估计值，S为总体标准差σ的估计值；$t_{\alpha/2,\nu}$为I型错误概率α相应的双侧 t 界值，那么总体均数μ的$100(1-\alpha)\%$置信区间（confidence interval, CI）为$\overline{X}\pm t_{\alpha/2,\nu}S/\sqrt{n}$。令$\delta$为容许误差，即为置信区间宽度的一半，那么有$\delta=t_{\alpha/2,\nu}S/\sqrt{n}$。经变换可得总体均数置信区间估计时，所需样本含量的计算公式为：

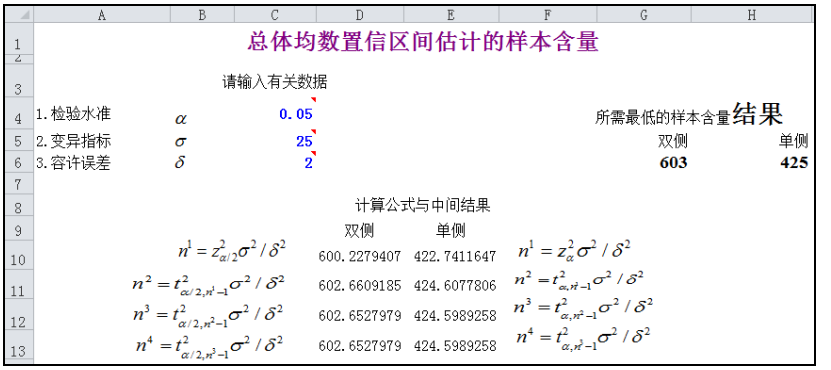
$$n=\left[\frac{t_{\alpha/2,\nu}S}{\delta}\right]^2 \qquad (14\text{-}8)$$

然而，$t_{\alpha/2,\nu}$与样本含量有关，样本含量减1是它们的自由度。在实际计算时，先用自由度$\nu=\infty$时的$t_{\alpha/2,\nu}$（此时的$t_{\alpha/2,\infty}=z_{\alpha/2}$）粗略估计样本含量$n^1$；第二次用$\nu=n^1-1$时的$t_{\alpha/2,\nu}$，代入公式（14-8）中粗略估计样本含量$n^2$；第三次用$\nu=n^2-1$时的$t_{\alpha/2,\nu}$代入公式（14-8）中粗略估计样本含量$n^3$；……；直至相邻两个样本含量结果趋于稳定为止。

例 14-11：成年男子血红蛋白的标准差为25g/l，在容许误差不超过2g/l，$\alpha=0.05$情况下，估计成年男子血红蛋白的置信区间需要调查多少病人？

解：该题是由样本均数估计总体均数的置信区间。打开光盘文件“14-6 置信区间样本含量.xlsx”的“总体均数区间估计”工作表，将已知数据输入到单元格C4、C5、C6（如图14-15

所示）中，得到双侧样本含量为 603，单侧样本含量为 425。在没有特别说明的情况下，一般选择双侧比较稳妥，所以估计成年男子血红蛋白的置信区间需要至少调查 603 个病人。



总体均数置信区间估计的样本含量

请输入有关数据

1. 检验水准	α	0.05
2. 变异指标	σ	25
3. 容许误差	δ	2

所需最低的样本含量结果

双侧	单侧
603	425

计算公式与中间结果

	双侧	单侧	
$n^1 = z_{\alpha/2}^2\sigma^2/\delta^2$	600.2279407	422.7411647	$n^1 = z_{\alpha}^2\sigma^2/\delta^2$
$n^2 = t_{\alpha/2,n^1-1}^2\sigma^2/\delta^2$	602.6609185	424.6077806	$n^2 = t_{\alpha,n^1-1}^2\sigma^2/\delta^2$
$n^3 = t_{\alpha/2,n^2-1}^2\sigma^2/\delta^2$	602.6527979	424.5989258	$n^3 = t_{\alpha,n^2-1}^2\sigma^2/\delta^2$
$n^4 = t_{\alpha/2,n^3-1}^2\sigma^2/\delta^2$	602.6527979	424.5989258	$n^4 = t_{\alpha,n^3-1}^2\sigma^2/\delta^2$

图 14-15　总体均数置信区间估计所需的样本含量

温馨提示：用户只需修改如图 14-15 所示的单元格区域 C4:C6 中的数据，便可获得总体均数置信区间估计所需的最低样本含量。

2. 总体率的置信区间

当总体率小于 0.2 或大于 0.8（极度偏离 0.5）时，总体率置信区间估计时所需样本含量的计算公式为：

$$n = \left[\frac{z_{\alpha/2}}{\sin^{-1}[\delta/\sqrt{\pi(1-\pi)}]}\right]^2 \qquad (14\text{-}9)$$

式中，n 为所需样本含量，π 为总体率或其估计值，$\sin(\bullet)$ 为反正弦函数，$z_{\alpha/2}$ 为 I 型错误概率 α 相应的双侧标准正态离差值，δ 为容许误差，即置信区间宽度的一半。

当总体率 $0.2 \leqslant \pi \leqslant 0.8$ 时，总体率置信区间估计时所需样本含量的计算公式为：

$$n = \left[z_{\alpha/2}/\delta\right]^2 \pi(1-\pi) \qquad (14\text{-}10)$$

式中符号含义与式（14-9）相同。

例 14-12：某病年复发率约为 10%，欲估计该病复发率的 95%置信区间，要求区间宽度的一半为 3%，问需要调查多少病人？

解：该题是由样本率估计总体率的置信区间。打开光盘文件“14-6 置信区间样本含量.xlsx”的“总体率区间估计”工作表，将已知数据输入到单元格 C4、C5、C6（如图 14-16 所示）中，得到双侧样本含量为 384，单侧样本含量为 270。在没有特别说明的情况下，一般选择双侧比较稳妥，所以该研究需要调查 384 个病人。

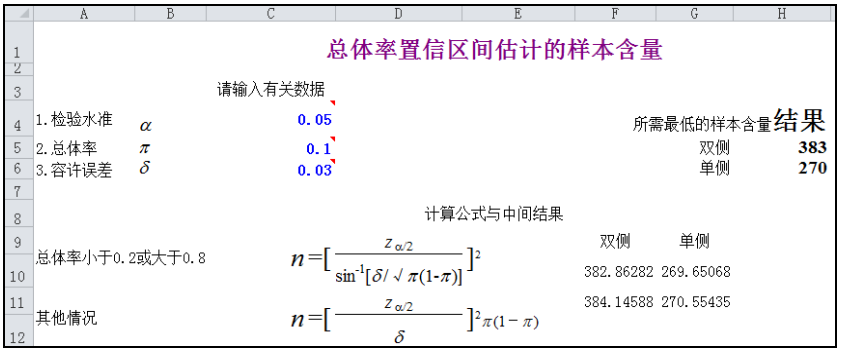


图 14-16 总体率置信区间估计所需的样本含量

温馨提示：用户只需修改如图 14-16 所示的单元格区域 C4:C6 中的数据，便可获得总体率置信区间估计所需的最低样本含量。

3. 有限总体率置信区间

有限总体率置信区间估计时所需样本含量的计算公式为：

$$n=\frac{z_{\alpha/2}^2\pi(1-\pi)N}{\delta^2(N-1)+z_{\alpha/2}^2\pi(1-\pi)} \tag{14-11}$$

式中，n为所需样本含量，N 为总体的大小，π 为总体率或其估计值，δ 为容许误差（即置信区间宽度的一半）、$z_{\alpha/2}$ 为 I 型错误概率 α 相应的双侧标准正态离差值。

例 14-13：从 8000 名生活在某地区的儿童中，随机抽取 100 名获得蛔虫患病率为 55%。问以容许误差为 6%，置信度为 95%进行估计，需用多大样本含量？

解：该题是由有限总体的样本率估计总体率的置信区间。打开光盘文件“14-6 置信区间样本含量.xlsx”的“有限总体率区间估计”工作表，将已知数据输入到单元格 C4、C5、C6、C7（如图 14-17 所示）中，得到双侧样本含量为 256，单侧样本含量为 182。在没有特别说明的情况下，一般选择双侧比较稳妥，所以该研究需要调查 256 名儿童。

温馨提示：用户只需修改如图 14-17 所示的单元格区域 C4:C7 中的数据，便可获得有限总体率置信区间估计所需的最低样本含量。

4. 分层抽样区间估计

按某种属性将研究总体分成若干层，如果已知每一层的人口数 N_i 和患病率 π_i，在 $100(1-\alpha)\%$ 置信度，容许误差为 δ 的情况下，研究所需样本含量的计算公式为：

$$n=\frac{z_{\alpha/2}^2\sum N_i^2\pi_i(1-\pi_i)/W_i}{N^2\delta^2+z_{\alpha/2}^2\sum N_i\pi_i(1-\pi_i)} \tag{14-12}$$

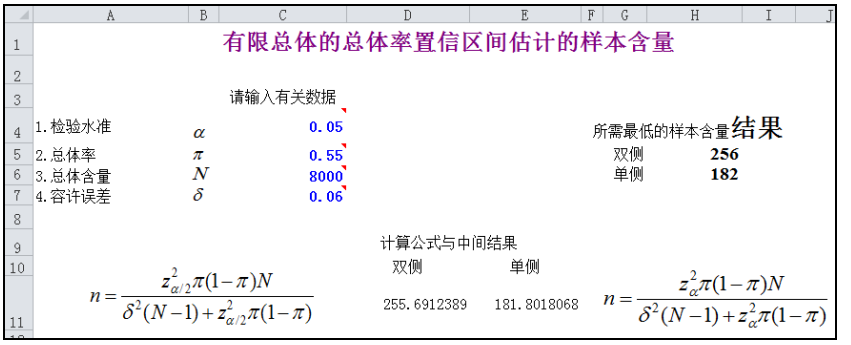


图 14-17　有限总体率置信区间估计所需的样本含量

其中 $N=\sum N_i$ 为总体大小，$W_i=N_i/N$ 为每一层的样本比例，$z_{\alpha/2}$ 为 I 型错误概率 α 相应的双侧标准正态离差值。

例 14-14：A、B、C、D、E 五个村各有 1880、2000、2140、2500、2600 人可供调查，据往年调查得知，五个村的血吸虫病患病率分别为 5%、8%、7%、15%、10%。在 95%置信度以及容许误差为 1%的情况下推断总体血吸虫病患病情况，各村需要调查多少人比较合理？

解：该题是通过分层抽样由样本率推断总体率的置信区间问题。打开光盘文件“14-6 置信区间样本含量.xlsx”的“分层抽样区间估计”工作表，将已知数据 $\alpha=0.05$、$\delta=0.01$ 分别输入到单元格 A10、B10 中，每层的患病率和总人数输入到单元格区域 D10:E59（最大可以输入 50 个层）中（如图 14-18 所示）。由此得到 A、B、C、D、E 五个村双侧样本含量分别为 421、448、480、560、583，该研究共计需要调查至少 2492 人。

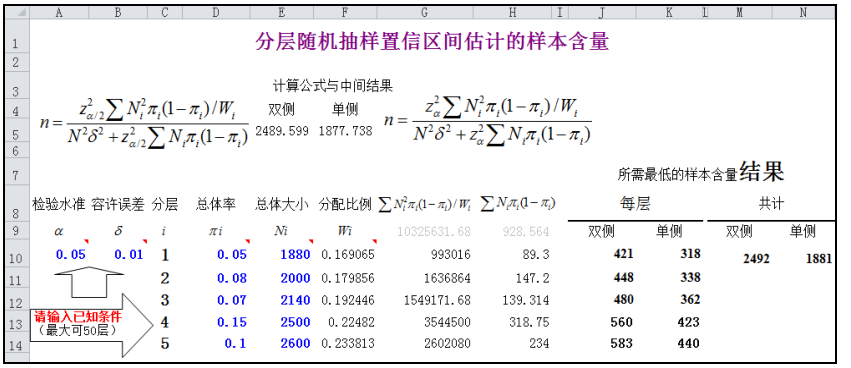


图 14-18　分层抽样置信区间估计所需的样本含量

温馨提示：用户只需修改如图 14-18 所示的单元格区域 A10:B10、D10:E59 中的数据，便可获得分层抽样置信区间估计所需的最低样本含量。该程序设定最大分层层数为 50。

14.3 检验功效的估计

前面已经提到，检验功效由犯Ⅱ型错误概率β的大小决定，即检验功效$=1-\beta$。$1-\beta=0.9=90\%$，表示所比较的两组确有差别时，100 次实验中平均有 90 次能发现差别。当假设检验出现“阴性”结果（如$P>0.05$）时，有两种可能：一是$1-\beta$较大，被比较的指标间很可能无差别；二是$1-\beta$较小，所比较的指标间很可能差别有意义，但由于样本含量不足而未能发现。因此，在估算样本含量时，要考虑检验效能；在试验结束时，也应利用研究例数计算检验效能，特别是当干预效果很小或结论不拒绝零假设时。

检验效能不得低于 75%，一般要求在 80%以上。在实验设计时，应尽量使各组样本例数相等，以取得较大的检验效能。

1. 定量数据单组设计与配对设计

由定量数据单组设计与配对设计样本含量的估计公式（14-1）可得：

$$t_{\beta,\nu}=\frac{\delta\sqrt{n}}{\sigma}-t_{\alpha/2,\nu} \qquad (14\text{-}13)$$

式中符号的意思也与公式（14-1）相同。值得注意的是：对于单组设计，n 是研究样本的样本含量，而对于配对设计 n 是研究样本的对子数；对于单组设计，δ 是样本均数对应的总体均数与给定的总体均数间差值，而对于配对设计 δ 是对子间差值的均数；对于单组设计，σ 是样本对应的总体标准差，而对于配对设计 σ 是总体差值标准差。通过 $t_{\beta,\nu}$ 可获得检验功效 $1-\beta$。

例 14-15：经连续非卧床腹膜透析治疗晚期糖尿肾病 6 例，测得生化指标钠的结果为：治疗前 138.33±2.28（$\bar{X}\pm S$），治疗后 139.67±1.41，差值 1.33±2.76。检验结果为：t=0.482,P>0.05。该结论是否可靠？（方亚，第 6 章样本含量估计，见董时富主编，生物统计学，科学出版社，2002）

解：该题是问配对 t 检验结果是否可靠的问题。打开光盘文件“14-7 检验功效.xlsx”的“定量数据单组与配对设计”工作表，将已知数据 $\alpha=0.05$、$n=6$、$\hat{\sigma}=S_d=2.76$、$\delta=1.33$ 分别依次输入到单元格区域 C4:C7 中，得到双侧检验功效只有 11.16%，单侧检验功效只有 22.10%（如图 14-19 所示）。说明检验效能太低，结论不可靠。

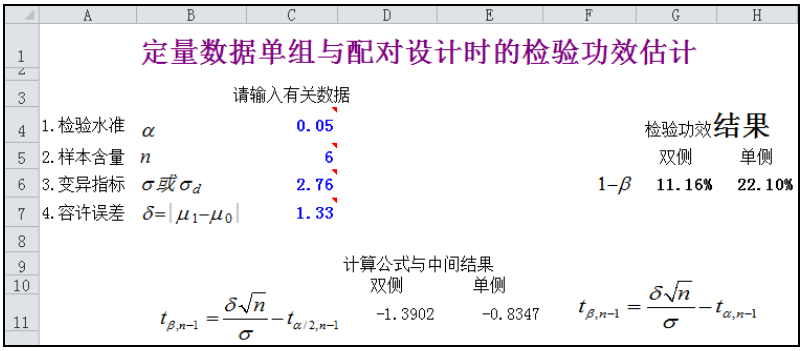


图 14-19　配对 t 检验的检验功效

温馨提示：用户只需修改如图 14-19 所示单元格区域 C4:C7 中的数据，便可获得样本均数与总体均数比较或配对均数比较的检验功效。

2. 定量数据成组设计

由两样本均数比较的 t 检验计算公式可得：

$$t_{\beta,n_1+n_2-2}=\frac{|\delta|}{\sigma\sqrt{1/n_1+1/n_2}}-t_{\alpha/2,n_1+n_2-2} \qquad (14\text{-}14)$$

式中符号的意思也与公式（14-2）相同。通过 t_{β,n_1+n_2-2} 可获得检验功效 $1-\beta$。

例 14-16：某降压药和安慰剂分别治疗两组高血压患者，服用 4 周后比较两组患者收缩压下降值（mmHg），测得结果 n_1=14，$\bar{X}_1$=10.38，S_1=6.32；n_2=16，$\bar{X}_2=6.62$，S_2=2.16。假设检验结果得到 t=2.34，P>0.05。问该研究的检验效能如何？

解：该题是问两样本 t 检验结果是否可靠。打开光盘文件“14-7 检验功效.xlsx”的“定量数据成组设计”工作表，将已知数据 $\alpha=0.05$、n_1=14、n_2=16 分别依次输入到单元格区域 C4:C6 中。S_1=6.32、S_2=2.16、$\bar{X}_1$=10.38、$\bar{X}_2=6.62$ 分别依次输入到单元格 B11、C11、F11、G11 中，在单元格 C7 中输入“=D11”，在单元格 C8 中输入“=H11”，得到双侧检验功效只有 57.52%，单侧检验功效只有 70.28%（如图 14-20 所示）。说明检验效能较低，结论不可靠。

温馨提示：用户只需修改如图 14-20 所示单元格区域 C4:C8 中的数据，以及单元格 B11、C11、F11、G11 中的均数标准差，便可获得两样本均数比较的检验功效。其中合并标准差的计算见 8.4.2 节公式。

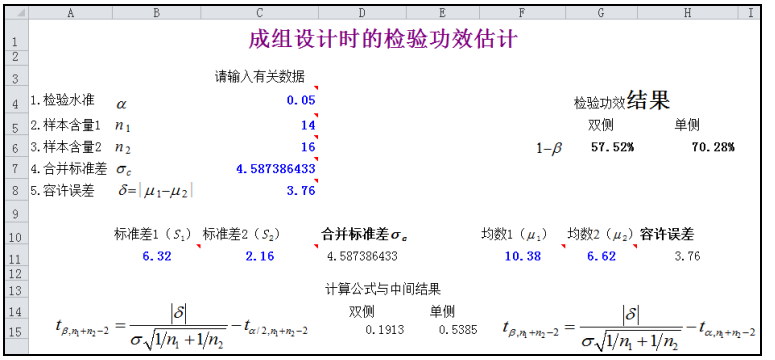


图 14-20 两样本均数 *t* 检验的检验功效

3. 定性数据配对设计

由样本含量估计公式（14-5）可得：

$$z_\beta=\frac{|\pi_{-+}-\pi_{+-}|\sqrt{n\pi_c}-z_{\alpha/2}\pi_c\sqrt{2}}{\sqrt{2\pi_{+-}\pi_{-+}}} \qquad (14\text{-}15)$$

式中符号的意思也与公式（14-5）相同。通过 z_β 可获得检验功效 $1-\beta$。

例 14-17：有 30 份标本接种于甲、乙两种培养基中，结果 π_{+-}=0.39，π_{-+}=0.48，检验得到 χ^2=0.24，P>0.05。结论为两种培养基培养结果无差异。问该结论是否可靠？

解：该题是问配对 χ^2 检验结果是否可靠。打开光盘文件“14-7 检验功效.xlsx”的“定性数据配对设计”工作表，将已知数据 $\alpha=0.05$、n=30、π_{+-}=0.39、π_{-+}=0.48 分别依次输入到单元格区域 C4:C7 中。得到双侧检验功效只有 7.50%，单侧检验功效只有 13.09%（如图 14-21 所示）。说明检验效能较低，结论不可靠。

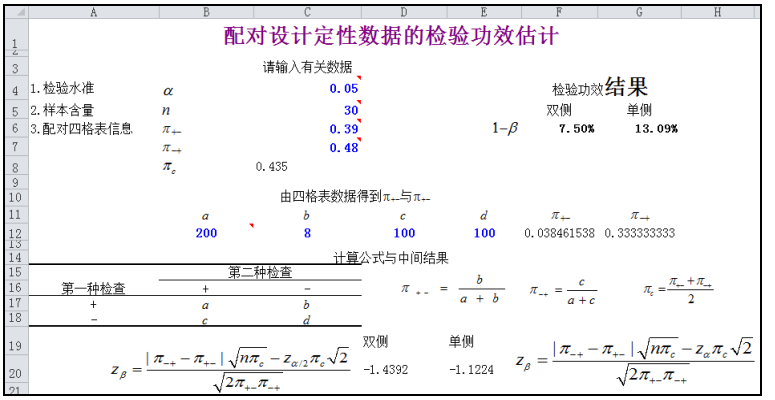


图 14-21 配对定性数据的检验功效

温馨提示：用户只需修改如图 14-21 所示单元格区域 C4:C7 中的数据，便可获得配对定性数据的检验功效。必要时可在单元格区域 B12:E12 中输入配对四格表数据，以便获得 π_{+-}、π_{-+}。

4. 定性数据成组设计例数相等

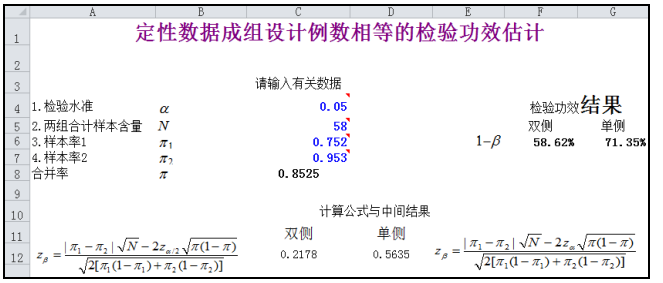
如果两样本含量相等，两组合计样本含量为 N，则有：

$$z_\beta = \frac{|\pi_1 - \pi_2|\sqrt{N} - 2z_{\alpha/2}\sqrt{\pi(1-\pi)}}{\sqrt{2[\pi_1(1-\pi_1)+\pi_2(1-\pi_2)]}} \tag{14-16}$$

这里的 π、π_1、π_2 与公式（14-6）中 p、p_1、p_2 的含义相同。通过 z_β 可获得检验功效 $1-\beta$。

例 14-18：为了进行两种药物疗效的比较，按入选条件选入 58 例，随机分为两组，每组 29 人。两组病人有效率分别为 75.2%和 95.3%，令 $\alpha = 0.05$，试估计检验功效。

解：该题是问两样本率比较结果是否可靠。打开光盘文件“14-7 检验功效.xlsx”的“定性数据成组设计例数相等”工作表，将已知数据 $\alpha = 0.05$、n =58、π_1 =0.752、π_2 =0.953 分别依次输入到单元格区域 C4:C7 中。得到双侧检验功效只有 58.62%，单侧检验功效只有 71.35%（如图 14-22 所示）。说明检验效能较低，结论不可靠。



定性数据成组设计例数相等的检验功效估计

请输入有关数据

1. 检验水准	α	0.05
2. 两组合计样本含量	N	58
3. 样本率1	π_1	0.752
4. 样本率2	π_2	0.953
合并率	π	0.8525

检验功效结果

	双侧	单侧
$1-\beta$	58.62%	71.35%

计算公式与中间结果

双侧	单侧
0.2178	0.5635

$z_\beta = \frac{|\pi_1-\pi_2|\sqrt{N}-2z_{\alpha/2}\sqrt{\pi(1-\pi)}}{\sqrt{2[\pi_1(1-\pi_1)+\pi_2(1-\pi_2)]}}$　　$z_\beta = \frac{|\pi_1-\pi_2|\sqrt{N}-2z_{\alpha}\sqrt{\pi(1-\pi)}}{\sqrt{2[\pi_1(1-\pi_1)+\pi_2(1-\pi_2)]}}$

图 14-22　定性数据成组设计例数相等的检验功效

温馨提示：用户只需修改如图 14-22 所示的单元格区域 C4:C7 中的数据，便可获得定性数据成组设计例数相等时的检验功效。

5. 定性数据成组设计例数不等

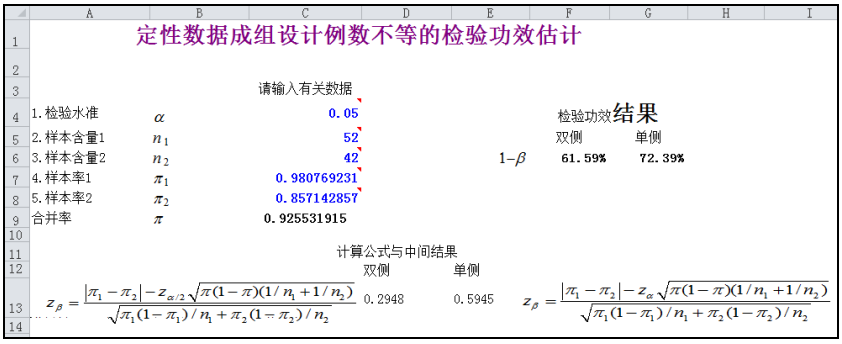
如果两样本含量不相等，两组分别为 n_1、n_2，则有

$$z_\beta = \frac{|\pi_1 - \pi_2| - z_{\alpha/2}\sqrt{\pi(1-\pi)(1/n_1+1/n_2)}}{\sqrt{\pi_1(1-\pi_1)/n_1+\pi_2(1-\pi_2)/n_2}} \tag{14-17}$$

这里的π、π_1、π_2与公式（14-6）中p、p_1、p_2的含义相同。通过z_β可获得检验功效$1-\beta$。

例 14-19：用异博停治疗试验组高血压病 52 人中 51 人有效，用硝苯吡啶治疗对照组高血压高血压病 42 人中 36 人有效，χ^2检验得到χ^2=3.51，$P>0.05$，问检验效能如何？

解：该题是问两样本率比较结果是否可靠，并且每组例数不相等。打开光盘文件“14-7 检验功效.xlsx”的“定性数据成组设计例数不等”工作表，将已知数据$\alpha=0.05$、n_1=52、n_2=42、π_1=51/52、π_2=36/42 分别依次输入到单元格区域 C4:C8 中。得到双侧检验功效只有 61.59%，单侧检验功效只有 72.39%（如图 14-23 所示）。说明检验效能较低，结论不可靠。



定性数据成组设计例数不等的检验功效估计

		请输入有关数据				
1. 检验水准	α	0.05			检验功效结果	
2. 样本含量1	n_1	52			双侧	单侧
3. 样本含量2	n_2	42		$1-\beta$	61.59%	72.39%
4. 样本率1	π_1	0.980769231				
5. 样本率2	π_2	0.857142857				
合并率	π	0.925531915				

计算公式与中间结果

双侧	单侧
0.2948	0.5945

$$z_\beta=\frac{|\pi_1-\pi_2|-z_{\alpha/2}\sqrt{\pi(1-\pi)(1/n_1+1/n_2)}}{\sqrt{\pi_1(1-\pi_1)/n_1+\pi_2(1-\pi_2)/n_2}}$$

$$z_\beta=\frac{|\pi_1-\pi_2|-z_{\alpha}\sqrt{\pi(1-\pi)(1/n_1+1/n_2)}}{\sqrt{\pi_1(1-\pi_1)/n_1+\pi_2(1-\pi_2)/n_2}}$$

图 14-23 定性数据成组设计例数不等的检验功效

温馨提示：用户只需修改如图 14-23 所示的单元格区域 C4:C8 中的数据，便可获得定性数据成组设计例数不等时的检验功效。

6. 简单线性相关分析

由直线相关的样本含量计算公式可得到：

$$z_\beta=\frac{1}{2}\left|\ln\left(\frac{1+\rho}{1-\rho}\right)\right|\sqrt{n-3}-z_{\alpha/2} \qquad (14\text{-}18)$$

这里所有符号与公式（14-7）中相同。通过z_β可获得检验功效$1-\beta$。

例 14-20：已知 n=15，样本相关系数 r=0.55，规定 α=0.05，问检验功效如何？

解：该题是问相关系数分析结果是否可靠。打开光盘文件“14-7 检验功效.xlsx”的“简单线性相关分析”工作表，将已知数据$\alpha=0.05$、n=15、r=0.55 分别依次输入到单元格区域 C4:C6 中。得到双侧检验功效只有 57.23%，单侧检验功效只有 69.05%（如图 14-24 所示）。说明检验效能较低，结论不可靠。

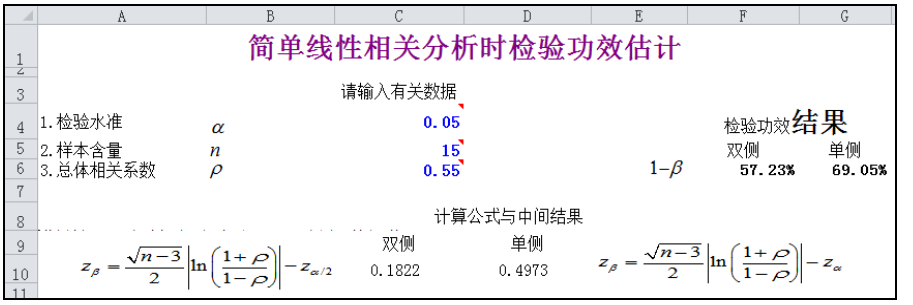


图 14-24　直线相关系数的检验功效

温馨提示：用户只需修改如图 14-24 所示的单元格区域 C4:C6 中的数据，便可获得线性相关系数的检验功效。

14.4　流行病学研究的样本含量与检验功效

流行病学的分析研究方法主要有病例对照研究和队列研究。病例对照研究是先分研究对象为病例组与对照组，然后回顾性追溯危险因素暴露情况；而队列研究恰好相反，是先分研究对象为暴露组与非暴露组，然后前瞻性追踪研究对象是否患病。

14.4.1　病例对照研究

在光盘文件“14-8 病例对照研究.xlsx”中有 5 个工作表，分别是 OR、成组样本量、配对样本量、成组检验功效、配对检验功效。

1. 优势比置信区间估计的样本含量

优势比（Odds Ratio，OR）的$100(1-\alpha)\%$置信区间为：

$$\widehat{OR}\times e^{\pm SE(\widehat{OR})} \tag{14-19}$$

假定病例组或对照组所需样本含量相等，均为 n，则估计该置信区间所需要的样本含量（Lemeshow S，et al，1988）为：

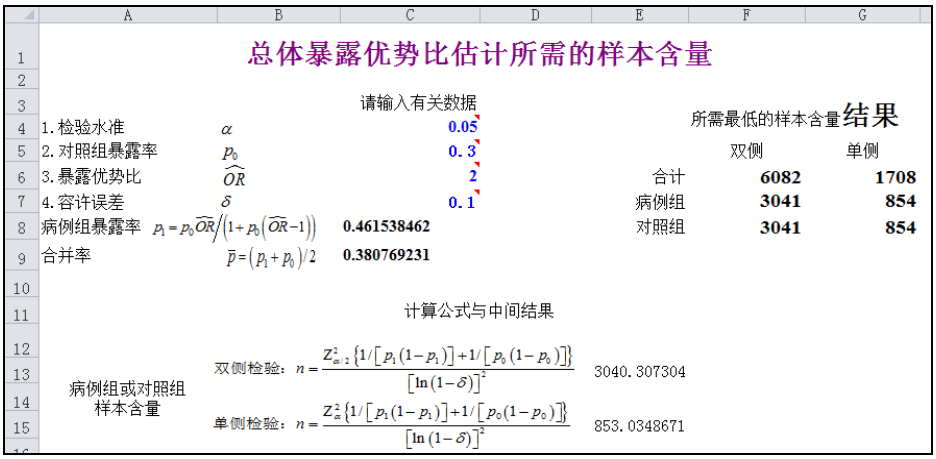
$$n=\frac{z_{\alpha/2}^2\left\{\dfrac{1}{p_1(1-p_1)}+\dfrac{1}{p_0(1-p_0)}\right\}}{\left[\ln(1-\delta)\right]^2} \tag{14-20}$$

式中，p_1、p_0 分别为病例组、对照组的暴露率，令置信区间宽度一半为：

$$\delta OR=e^{\ln(OR)}-e^{\ln(OR)-z_{\alpha/2}SE(\widehat{OR})},\quad \delta=1-e^{-z_{\alpha/2}SE(\ln(\widehat{OR}))}$$

例 14-21：假定某研究的优势比 OR=2，对照组暴露率 p_0=0.30，在 10%范围内，以 95%置信区间估计总体优势比，问最少需要多少样本含量？

解：打开光盘文件“14-8 病例对照研究.xlsx”的“OR”工作表，将已知数据 $\alpha=0.05$、p_0=0.30、OR=2、$\delta=0.1$ 分别依次输入到单元格区域 C4:C7 中。得到病例组、对照组各需要 3041，两组共需要 6082；如果采用单侧，则病例组、对照组各需要 854，两组共需要 1708（如图 14-25 所示）。



总体暴露优势比估计所需的样本含量

		请输入有关数据			所需最低的样本含量结果	
1. 检验水准	α	0.05			双侧	单侧
2. 对照组暴露率	p_0	0.3		合计	6082	1708
3. 暴露优势比	$\widehat{OR}$	2		病例组	3041	854
4. 容许误差	δ	0.1		对照组	3041	854
病例组暴露率	$p_1=p_0\widehat{OR}/(1+p_0(\widehat{OR}-1))$	0.461538462				
合并率	$\bar{p}=(p_1+p_0)/2$	0.380769231				

计算公式与中间结果

病例组或对照组样本含量	双侧检验：$n=\frac{Z_{\alpha/2}^2\{1/[p_1(1-p_1)]+1/[p_0(1-p_0)]\}}{[\ln(1-\delta)]^2}$	3040.307304
	单侧检验：$n=\frac{Z_{\alpha}^2\{1/[p_1(1-p_1)]+1/[p_0(1-p_0)]\}}{[\ln(1-\delta)]^2}$	853.0348671

图 14-25 优势比置信区间估计的样本含量

2. 非配比设计优势比假设检验的样本含量

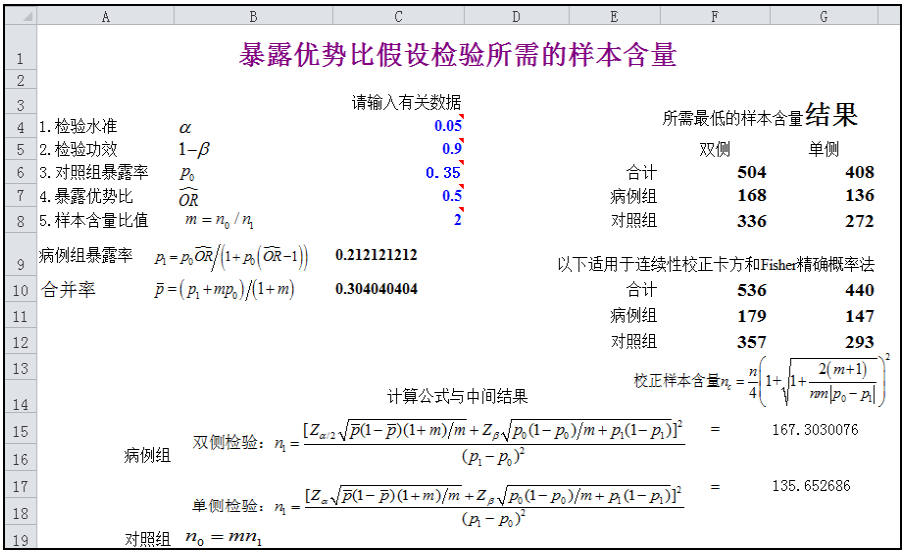
优势比假设检验的零假设 H_0：$OR=1$，病例组所需样本含量的计算公式：

$$n_1=\frac{[z_{\alpha/2}\sqrt{\bar{p}(1-\bar{p})(1+m)/m}+z_{\beta}\sqrt{p_0(1-p_0)/m+p_1(1-p_1)}]^2}{(p_1-p_0)^2} \qquad (14\text{-}21)$$

式中，p_1、p_0 分别为病例组、对照组的暴露率，$m=n_0/n_1$ 即对照组与病例组样本含量之比，平均率 $\bar{p}=(p_1+mp_0)/(1+m)$。如果给定优势比，则 $p_1=p_0\widehat{OR}/\left(1+p_0\left(\widehat{OR}-1\right)\right)$。

例 14-22：假定某研究的优势比 OR=0.5，对照组暴露率 p_0=0.35，对照组样本量为病例组的两倍，检验水准 α=0.05，在问 90%检验功效情况下需要多少样本含量？

解：打开光盘文件“14-8 病例对照研究.xlsx”的“成组样本量”工作表，将已知数据 $\alpha=0.05$、$1-\beta$=0.9、p_0=0.35、OR=0.5、$m=2$ 分别依次输入到单元格区域 C4:C8 中。得到病例组、对照组各需要 168、336，两组共需要 504；如果采用单侧，则病例组、对照组各需要 136、272，两组共需要 408（如图 14-26 所示）。



暴露优势比假设检验所需的样本含量

请输入有关数据

1. 检验水准	α	0.05
2. 检验功效	$1-\beta$	0.9
3. 对照组暴露率	p_0	0. 35
4. 暴露优势比	$\widehat{OR}$	0.5
5. 样本含量比值	$m=n_0/n_1$	2
病例组暴露率	$p_1=p_0\widehat{OR}/(1+p_0(\widehat{OR}-1))$	0.212121212
合并率	$\bar{p}=(p_1+mp_0)/(1+m)$	0.304040404

所需最低的样本含量结果

	双侧	单侧
合计	504	408
病例组	168	136
对照组	336	272

以下适用于连续性校正卡方和Fisher精确概率法

合计	536	440
病例组	179	147
对照组	357	293

校正样本含量 $n_c=\frac{n}{4}\left(1+\sqrt{1+\frac{2(m+1)}{nm|p_0-p_1|}}\right)^2$

计算公式与中间结果

病例组 双侧检验：$n_1=\frac{[Z_{\alpha/2}\sqrt{\bar{p}(1-\bar{p})(1+m)/m}+Z_{\beta}\sqrt{p_0(1-p_0)/m+p_1(1-p_1)}]^2}{(p_1-p_0)^2}$ = 167.3030076

单侧检验：$n_1=\frac{[Z_{\alpha}\sqrt{\bar{p}(1-\bar{p})(1+m)/m}+Z_{\beta}\sqrt{p_0(1-p_0)/m+p_1(1-p_1)}]^2}{(p_1-p_0)^2}$ = 135.652686

对照组 $n_0=mn_1$

图 14-26　非配比设计优势比假设检验的样本含量

3. 配比设计优势比假设检验的样本含量

对于配比设计，优势比假设检验的零假设 H_0：$OR=1$，病例组所需样本含量的计算公式为：

$$n_1=\frac{\left(\frac{\sigma_1}{\sigma_{\widehat{OR}}}z_{\alpha/2}+z_{\beta}\right)^2}{\delta^2},\quad \sigma_{\widehat{OR}}=\sqrt{\sum_{k=1}^{m}\frac{kt_k\widehat{OR}(m-k+1)}{\left(k\widehat{OR}+m-k+1\right)^2}}$$

$$\sigma_1=\sqrt{\sum_{k=1}^{m}\frac{kt_k(m-k+1)}{(k+m-k+1)^2}},\quad \delta=\frac{\left(f(\widehat{OR})-f(1)\right)}{\sigma_{\widehat{OR}}} \qquad (14\text{-}22)$$

式中：

$$f(1)=\sum_{k=1}^{m}\frac{kt_k}{k+m-k+1},\quad f\left(\widehat{OR}\right)=\sum_{k=1}^{m}\frac{kt_k\widehat{OR}}{k\widehat{OR}+m-k+1}$$

其他与前面相同（Dupont WD，1988，1990）。

例 14-23：假定某研究的优势比 OR=0.5，对照组暴露率 p_0=0.35，病例组与对照组相关系数为 0.2，对照组样本量为病例组的 5 倍，检验水准 α=0.05，问在 90%检验功效情况下需要多少样本含量？

解：打开光盘文件“14-8 病例对照研究.xlsx”的“配比样本量”工作表，将已知数据

$\alpha=0.05$、$1-\beta=0.9$、$p_0=0.35$、$OR=0.5$、$m=2$、$r=0.2$ 分别依次输入到单元格区域 C4:C9 中。得到病例组、对照组各需要 191、382，两组共需要 573；如果采用单侧，则病例组、对照组各需要 155、310，两组共需要 465。（如图 14-27 所示）。

暴露优势比假设检验所需的样本含量

			请输入有关数据
1. 检验水准	α		0.05
2. 检验功效	$1-\beta$		0.9
3. 对照组暴露率	p_0		0.35
4. 暴露优势比	$\widehat{OR}$		0.5
5. 样本含量比值	$m=n_0/n_1$		2
6. 相关系数	r		0.2
病例组暴露率	$p_1=p_0\widehat{OR}\big/\left(1+p_0\left(\widehat{OR}-1\right)\right)$		0.212121212
	$p_{0r}=\left(p_1p_0+r\sqrt{p_1(1-p_1)p_0(1-p_0)}\right)\big/p_1$		0.533847763
	$p_{0r}=\left(p_0(1-p_1)-r\sqrt{p_1(1-p_1)p_0(1-p_0)}\right)\big/(1-p_1)$		0.300502525

所需最低的样本含量结果

	双侧	单侧
合计	573	465
病例组	191	155
对照组	382	310

计算公式与中间结果

k	t_k	σ_{OR}	σ_1	$f(1)$	$f(OR)$	δ	双侧	单侧
1	0.377318932	0.060371029	0.0838	0.125773	0.075463786	-0.246680762	190.916912	154.5313834
2	0.176721411	0.044180353	0.0393	0.117814	0.088360706			

图 14-27 配比设计优势比假设检验的样本含量

成组检验功效、配对检验功效的计算公式可由其对应的样本含量计算公式简单推导获得，其应用简单，在此不再赘述。

14.4.2 队列研究

在光盘文件“14-9 队列研究.xlsx”中也有 5 个工作表，分别是 RR、成组样本量、配对样本量、成组检验功效、配对检验功效。其应用与病例对照研究十分相似，在此只列举相对危险度 RR 的置信区间估计所需样本含量，其他不再赘述。

相对危险度（Relative Risk，RR）的$100(1-\alpha)\%$置信区间为：

$$\widehat{RR}\times e^{\pm SE(\widehat{RR})} \tag{14-23}$$

假定病例组或对照组所需样本含量相等，均为 n，则估计该置信区间所需要的样本含量（Lemeshow S，et al，1988）为：

$$n=\frac{z_{\alpha/2}^2\left\{\dfrac{1-p_1}{p_1}+\dfrac{1-p_0}{p_0}\right\}}{\left[\ln\left(1-\delta\right)\right]^2} \tag{14-24}$$

式中，p_1、p_0 分别为病例组、对照组的暴露率，令置信区间宽度一半为：

$$\delta\widehat{RR}=e^{\ln(\widehat{RR})}-e^{\ln(\widehat{RR})-z_{\alpha/2}SE(\widehat{RR})},\quad \delta=1-e^{-z_{\alpha/2}SE(\ln(\widehat{RR}))}$$

例 14-24：假定某研究的相对危险度 RR=2，对照组暴露率 p_0=0.30，在 10%范围内，以 95%置信区间估计总体优势比，问最少需要多少样本含量？

解：打开光盘文件“14-9 队列研究.xlsx”的“RR”工作表，将已知数据 $\alpha=0.05$、p_0=0.30、RR=2、$\delta=0.1$ 分别依次输入到单元格区域 C4:C7 中。得到病例组、对照组各需要 1039，两组共需要 2078；如果采用单侧，则病例组、对照组各需要 732，两组共需要 1464（如图 14-28 所示）。

	A	B	C	D	E	F	G
1	总体相对危险度估计所需的样本含量						
2							
3			请输入有关数据				
4	1. 检验水准	α	0.05			所需最低的样本含量结果	
5	2. 非暴露组发病率	p_0	0.3			双侧	单侧
6	3. 相对危险度	$\widehat{RR}$	2		合计	2078	1464
7	4. 容许误差	δ	0.1		非暴露组	1039	732
8	暴露组发病率	$p_1=\widehat{RR}\times p_0$	0.6		暴露组	1039	732
9	合并率	$p=(p_1+p_0)/2$	0.45				
10							
11			计算公式与中间结果				
12		双侧检验：$n=\frac{z_{\alpha/2}^2\left[(1-p_1)/p_1+(1-p_0)/p_0\right]}{\left[\ln(1-\delta)\right]^2}=$			1038.153713		
13	非暴露组或暴露组样本含量						
14		单侧检验：$n=\frac{z_{\alpha}^2\left[(1-p_1)/p_1+(1-p_0)/p_0\right]}{\left[\ln(1-\delta)\right]^2}=$			731.1727432		
15							
16							

图 14-28　优势比置信区间估计的样本含量

14.5　方差分析等方法的样本含量与检验功效

14.5.1　方差分析的检验功效

Mick O’Neill 和 Peter C. Thomson 给出的方差分析检验功效：

$$1-\beta=P(F>f)=1-\sum_{r=0}^{\infty}P\left(r;\frac{\lambda}{2}\right)G\left(\frac{\frac{\nu_1}{\nu_2}f}{1+\frac{\nu_1}{\nu_2}f};\frac{\nu_1}{2}+r,\frac{\nu_2}{2}\right)\quad(14\text{-}25)$$

其中 $f=F_{临界值}$，由检验水准 α、分子自由度 $\nu_1=df_1=k-1$、分母自由度 $\nu_2=df_2$ 决定。λ 为非中心参数，$P\left(r;\frac{\lambda}{2}\right)$ 为 Poisson 变量概率，$G(x;a,b)$ 为 beta I 型函数。

例 14-25：某研究者采用完全随机设计方差分析进行假设检验，所分析数据有 4 个

处理组，每组 8 个个体，每组标准差约为 1，拟在$\alpha=0.05$检验水准下发现δ=2 的差异，问检验效能如何？

解：打开光盘文件“14-10 方差分析的检验功效.xlsx”，将已知数据$\alpha=0.05$、k=4、n_i=8、S_i^2=1、δ=2 分别依次输入到单元格区域 C4:C8 中。得到检验功效为 89.36%（如图 14-29 所示），说明检验效能尚可，结论较可靠。

温馨提示：用户只需修改如图 14-29 所示的单元格区域 C4:C8 中的数据，便可获得方差分析的检验功效。该程序包括了宏函数，鉴于 Office 2010 中 Excel 存在宏安全级，初次打开此程序文件时在弹出的“启用宏”提示框中应单击“确定”按钮。此程序适用于完全随机设计、随机区组设计、拉丁方设计的样本含量和检验功效估计，在“设计类型”右侧单元格（见单元格 C3）中输入或选择 0、1、2 分别对应完全随机设计、随机区组设计、拉丁方设计。

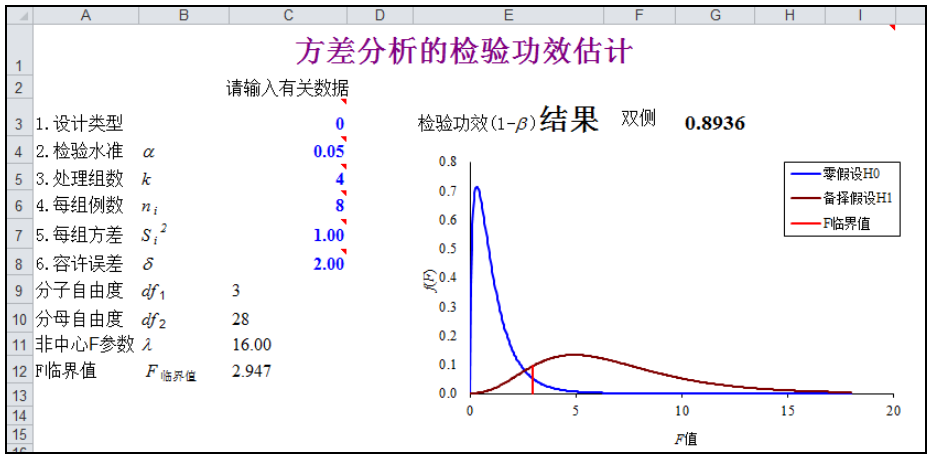


图 14-29 方差分析的检验功效

14.5.2 方差分析的样本含量

采用功能区中的“数据→数据工具→模拟分析→单变量求解”命令，可由光盘文件“14-10 方差分析的检验功效.xlsx”获得一定检验功效下的样本含量。比如打算在检验功效为 95%的情况下获得例 14-25 的每组例数n_i，可单击功能区“数据→数据工具→模拟分析→单变量求解”命令，在弹出的单变量求解对话框中，“目标单元格”选取“G3”，“目标值”输入“0.95”，可变单元格选取“C6”，单击“确定”按钮后，获得检验功效为 95%（实际当前解为 94.94%）时$n_i=10$（见单元格 C6）。按同样方法可以估计处理组数k。

14.5.3 其他方法的样本含量与检验功效

实际上，利用第 11 章介绍的单变量求解或规划求解（后者需要取消保护，利用密码为“0”进行保护撤销），可由光盘文件“14-5 假设检验样本含量.xlsx”的每一个工作表（在样本含量不太低的情况下）的样本含量中间结果（公式中包含检验功效单元格）获得检验功效；同样，由光盘文件“14-7 检验功效.xlsx”的每一个工作表（在样本功效不太低的情况下）的检验功效获得样本含量。

一般线性回归分析时，样本含量可以根据变量个数进行粗略估计，一般认为，观察个体数 n 与变量个数 p 的比例一般至少应为 $n:p=5\sim10$，即样本含量应至少是变量个数的 5～10 倍；和一般线性回归不同，logistic 回归采用最大似然估计（Maximum Likelihood Estimation，MLE）获得参数估计值，而不是一般最小二乘法，MLE 依赖于大样本渐近正态性质，在样本含量较少的情况下，很可能导致参数估计不收敛。一般认为，logistic 回归分析的观察个体数 n 与变量个数 p 的比例一般至少应为 $n:p=15\sim20$，即样本含量至少是变量个数的 15~20 倍。Cox 比例风险模型的样本含量可参照 logistic 回归。

为了较准确估计样本含量，读者可参见一般线性回归所需样本含量计算网址：http://www.danielsoper.com/statcalc/calc01.aspx 或 logistic 回归所需样本含量计算网址 http://www.dartmouth.edu/~eugened/power-samplesize.php。

第 15 章

高级科研统计图与 Meta 分析

本章将在第 6 章统计图表的基础上介绍统计地图、正态 P-P 图与 Q-Q 图、生存曲线图、重复测量数据的曲线图、Meta 分析及其置信区间图与漏斗图。其他图形读者可由此举一反三获得。

15.1 统计地图

统计地图（statistical map）用于表示某现象的数量在地域上的分布。绘制时首先应绘制一张空白地图，然后根据指标的大小分成不同的等级，采用不同的线条或不同的颜色绘制在地图上，以图例说明不同线条或不同颜色的意义。

创建空白地图是制作统计地图的关键，下面将采用 Excel 绘图工具制作统计地图。

选择功能区“插入→插图→形状”命令，任意选择一个形状并在绘图区绘制后，功能区中就会出现“绘图工具”选项，单击下方的“格式”，如图 15-1 所示。

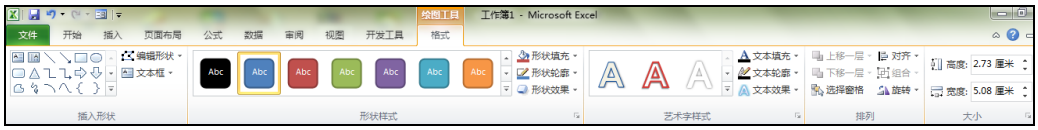


图 15-1　绘图工具的格式

例 15-1：经调查，某省 8 个行政地区各自的恶性肿瘤死亡专率（1/10 万）有所不同，其中地区 3、5、7 的恶性肿瘤死亡专率在 60/10 万以下，地区 2、6、8 为 60～80/10 万，地区 1、4 在 80/10 万以上，试绘制统计地图（如图 15-2 所示）表示该资料。

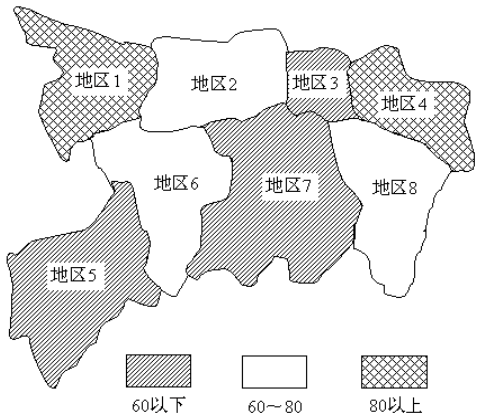


图 15-2　某省 8 个地区恶性肿瘤死亡专率（1/10 万）

解：为了绘制如图 15-2 所示的这个图形，可按下列步骤操作。

1. 绘制空白地图

① 绘制空白地图轮廓。单击“插入→插图→形状→线条→任意多边形”命令。此时箭头光标变为“+”，将光标移至需要绘图的地方，按住鼠标左键后，“+”变成画笔，仿照图

15-2 绘制地区 1 的多边形；按同样方法绘制其他 7 个地区的多边形。绘制多边形时，起点可从任意地方开始，中途可以松开鼠标左键，然后再按住鼠标左键继续绘制，快速双击鼠标左键可结束绘制。绘制的空白地图轮廓如图 15-3 所示（见光盘文件“15-1 绘图工具绘制统计地图.xlsx”）。

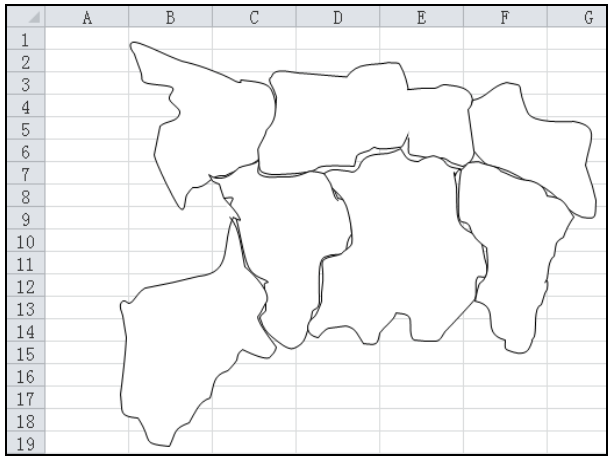


图 15-3 由“绘图”工具绘制的某省空白地图轮廓

② 编辑空白地图轮廓。8 个多边形应逐个编辑。单击选择 8 个多边形中的一个（如地区 1），单击鼠标右键，在弹出的快捷功能区中选择“编辑顶点”，此时多边形上显示顶点句柄，通过拖动这些句柄可对曲线进行编辑。在出现句柄的情况下，将鼠标移至一个句柄上，单击鼠标右键，可对该句柄进行编辑，如删除、添加、平滑顶点等（如图 15-4 所示）。如果顶点太多，可通过“绘图工具→格式→排列→上/下移一层”进行辅助编辑，经过如此编辑后的空白地图如图 15-5 所示。

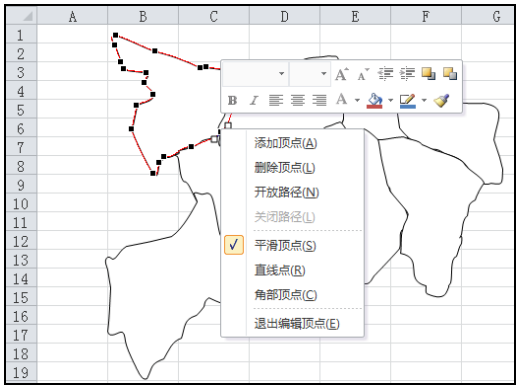


图 15-4 多边形顶点的编辑

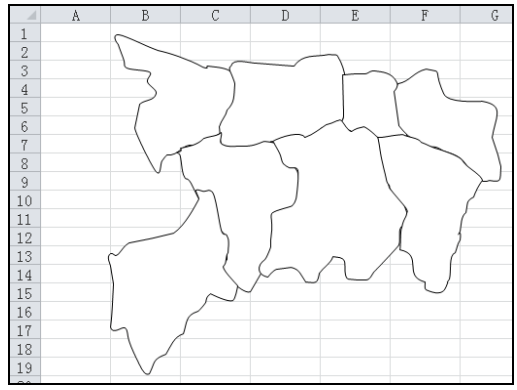


图 15-5 编辑后的空白地图

2. 在空白地图中添加线条或颜色

可对 8 个多边形逐个添加线条或颜色。选取地区 1，单击鼠标右键，在弹出的快捷功能区中选择“设置形状格式→填充→图案填充”，前景颜色选黑色，背景颜色选白色，选中第 5 行最后一列的图案，单击“确定”按钮。

其他 7 个地区按照图 15-2 以及以上操作方法，添加线条。添加颜色可仿照此操作。

3. 添加图例与文本框标签

选择功能区“插入→插图→形状→矩形”命令，在地图下方适当位置画一个矩形。在该矩形上单击鼠标右键，在弹出的快捷功能区中选择“设置形状格式→填充→图案填充”，前景颜色选黑色，背景颜色选白色，选中第 5 行最后一列的图案（即与所在的多边形区域的填充效果一致），单击“确定”按钮。单击“绘图工具”中的文本框，在矩形下方单击鼠标左键，向右下方拖动光标，获得适当大小的文本框，选择“横排文本框”，在其中键入“60 以下”。在文本框周围用鼠标右键单击文本框，选择“设置形状格式→线条颜色”，选择无线条颜色。

其他 2 个图例以及 8 个地区的文本框标签，仿照上述方法设置。

4. 组合统计地图

以上绘制出的统计图，有 8 个多边形、8 个地区的文本框标签、3 个图例、3 个图例的文本框标签，这 22 个部分“各自为阵”可分别编辑。如果要复制到 Microsoft Word 等文本编辑文档中，需复制 22 次不说，各个部分的相对位置也很难保持不变，为此需要把它们组合成一个图形。

按住 Shift 键不放，用鼠标单击选取需要组合的对象，选取后的对象四周将出现 8 个白色句柄。选取完所有对象（共 22 个）后，在其中一个对象上单击鼠标右键，在快捷功能区中选择“组合→组合”命令。

组合后的图形，还可以进行“重新组合”或“取消组合”。可以还原成组合前的状态。

可将组合后的图形复制，在 Word 中选择性粘贴“图片”，适当调整图形大小，即可完成如图 15-2 所示的统计地图的绘制。

绘图工具小技巧：在使用“绘图”工具时，使用 Shift 键+直线（或箭头）可绘制线条为水平线、垂直线或角度为 15° 倍数的直线；使用 Shift 键+矩形可得正方形；使用 Shift 键+椭圆可得圆形；使用 Shift 键+自选图形可获得大约对称的自选图形。目前电子地图（包括纸质扫描地图）很多，网上通过 Google 等搜索引擎很容易获得，为制作空白统计图提供了很多方便。另外，出于美观考虑，图形背景的网格线可以在“视图”中的“显示”命令组框内勾选“网格线”进行消除。

Excel 2003 以前版本，如 Excel 2000 中集成了微软地图插件，帮助用户根据数据制作各个国家、各省地图。但为了更大利益，微软已不再在 Excel 2003 以上版本支持 Excel 微

软地图，而是从 Excel 中剥离出来，开发出了 MapPoint 软件，且仅在北美和欧洲两个地区销售。一些精通宏制作的网友经过努力，正试图开发出能够在 Excel 中制作层析地图的宏。

15.2 P-P 图与 Q-Q 图

P-P 图与 Q-Q 图可以用来直观判断一个变量的分布是否服从某一给定的分布（如正态分布）。

以一个变量的观察累计概率（observed cumulative probability）为横坐标，按照某一检验分布（如正态、*t*、卡方等分布）获得的期望累计概率（expected cumulative probability）为纵坐标绘制散点图，如果散点分布呈一条直线，则认为该变量数据服从给定的分布。这样获得的散点图称为 P-P 图（P-P plot）。其中 P 代表概率（probability），如果检验分布为正态分布，则绘出的图形为正态 P-P 图。

以一个变量的观察值（observed value）为横坐标，按照某一检验分布（如正态、*t*、卡方等分布）获得的期望分位数（expected quantile）为纵坐标绘制散点图，如果散点分布呈一条直线，则认为该变量数据服从给定的分布。这样获得的散点图称为 Q-Q 图（Q-Q plot）。其中 Q 代表分位数（quantile），如果检验分布为正态分布，则绘出的图形为正态 Q-Q 图。

例 15-2：*以第 7 章例 7-1 的 130 例红细胞计数数据为例，绘制正态 P-P 图与正态 Q-Q 图。*

解：该例对应的光盘文件为“15-2 P-P 与 Q-Q 图.xlsx”。绘制正态 P-P 图与正态 Q-Q 图按如下步骤进行。

15.2.1 创建绘图数据

①将已知的 130 例红细胞计数数据键入到单元格区域 A3:A132 中，在单元格 A2 中键入“0”备用。

②在单元格 G2 中输入公式“=AVERAGE($A$3:$A$132)”，在单元格 H2 中输入公式“=STDEV($A$3:$A$132)”，分别产生 130 个数据的均数、标准差。

③在单元格 C3 中输入公式“=NORMDIST(A3,$G$2,$H$2,1)”，求红细胞计数为 4.54（A3）的正态分布的累积概率。

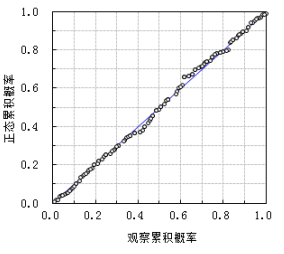
④在单元格 D3 中输入公式“=PERCENTRANK($A$2:$A$132,A3,6)”，求红细胞计数为 4.54（A3）在单元格区域 A2:A132（0 加上 130 个红细胞计数）中所排列的位次，PERCENTRANK 函数后面的 6 表示小数位数精确到小数点后 6 位。假如有 100 例数据，因为 PERCENTRANK 函数的排列位次是从 0 开始的，然后 1、2、……、100，所以为了正确排位还应在整个数据中增加一个比最小值还小的数。这样，0 的排列位次仍为 0，最小值

排列位次为“1/130= 0.007692”……，第 14 个值排列位次为“14/130 = 0.107692”……，最大值排列位次为“130/130=1”。因为它的取值在 0～1 之间，为了与正态累积概率相对应，不妨叫它“观察累积概率”。

⑤在单元格 E3 中输入公式“=NORMINV(D3,$G$2,$H$2)”，求均数、标准差分别为 4.726615（G2）、0.598873（H2）的正态分布中，观察累积概率为“0.423076（D3）=55/130”对应的正态分位数。

⑥选取单元格区域 C3:E3，将鼠标移向右下角的填充柄处，按住鼠标向下拖放填充 3 个公式至单元格 E132，并调整这三列的小数位数至 6 位（如图 15-6 所示）。

	A	C	D	E	G	H
1	红细胞计数	正态累积概率	观察累积概率	正态分位数	均数	标准差
2	0.00				4.726615	0.598873
3	4.54	0.377668	0.423076	4.610416		
4	3.97	0.103223	0.107692	3.984674		
5	4.85	0.581615	0.584615	4.854603		
6	5.45	0.886459	0.884615	5.4442		
7	5.41	0.873090	0.869230	5.3990		
8	5.72	0.951418	0.946153	5.6899		
9	3.44	0.015841	0.015384	3.4330		
10	4.57	0.396847	0.430769	4.6221		
11	5.35	0.851046	0.846153	5.3375		
12	4.08	0.140133	0.138461	4.0754		
13	4.31	0.243319	0.238461	4.3006		
14	3.73	0.048041	0.061538	3.8031		
15	3.64	0.034805	0.030769	3.6069		
16	4.14	0.163659	0.161538	4.1348		
17	4.26	0.217944	0.215384	4.2547		
18	3.48	0.018689	0.023076	3.5324		
19	4.38	0.281369	0.292307	4.3992		
20	4.48	0.340243	0.346153	4.4896		



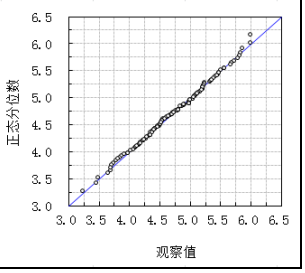


图 15-6　创建正态 P-P 图与正态 Q-Q 图的数据

15.2.2　绘制正态 P-P 图

①选择“插入→图表→*XY* 散点图”命令（第 1 个子图表类型）。

②在弹出的图表框内单击鼠标右键选择“选择数据”。弹出“选择数据源”对话框，删除所有系列，单击“添加”按钮。在“*X* 轴系列值”右侧用鼠标选取单元格区域 D3:D132，在“*Y* 轴系列值”右侧用鼠标选取单元格区域 C3:C132，单击“确定”按钮。

③选择功能区“图表工具→布局→标签→坐标轴标题”对话框中，在主要横/纵坐标轴标题内分别填入“观察累积概率”、“正态累积概率”；在“图表工具→布局→坐标轴→网格线→主要横/纵网格线”命令中，*X* 轴与 *Y* 轴均选择“次要网格线”。

④在横坐标上单击鼠标右键，选择“设置坐标轴格式→坐标轴选项”，主要刻度线类型与次要刻度线类型均选中“内部”，“最小值、最大值、主要刻度单位、次要刻度单位”均选“固定”，值分别为 0、1、0.2、0.1；选择“数字”，在分类下拉列表中选择数值，小数位数设置为 1 位。单击“关闭”按钮。

按同样方法设置纵坐标的图案、刻度、数字。

⑤在 *Y* 轴次要网格线上单击鼠标右键，选择“设置次要网格线格式”，“线条颜色”为实线)，颜色为“25%灰色”（第 4 行最后一种颜色)，“线型”→“宽度”为 0.25 磅，“短划线类型”选虚线(第三条)，单击“关闭”按钮。

以相同方式设置 *X* 轴次要网格线。

⑥在数据系列（散点）标志上单击鼠标右键，选择“设置数据系列格式→数据标记选项”数据标记类型选为内置，类型选择圆圈，大小改为 3 磅，圆圈颜色可以在“标记线颜色”中更改。单击“关闭”按钮。

⑦适当调整绘图区的大小，调整整个图表区域的字体大小等，可获得如图 15-6 中间所示的正态 P-P 图。

15.2.3 绘制正态 Q-Q 图

绘制正态 Q-Q 图有两种方法：一是仿照上述绘制正态 P-P 图的方法进行绘制；二是一份绘制好的正态 P-P 图，适当修改有关数据来源、横纵坐标轴及坐标轴的标题。第二种方法更方便快捷，下面介绍第二种方法的具体绘制步骤。

①在图表区选中绘制好的正态 P-P 图，复制、粘贴到同一张工作表上。

②在所复制图形的图表区或绘图区单击鼠标右键，选择“选择数据”。弹出“选择数据源”对话框，删除所有系列，单击“添加”按钮。在“*X* 值”右侧用鼠标选取单元格区域 A3:A132，在“*Y* 值”右侧用鼠标选取单元格区域 E3:E132，单击“确定”按钮。

③在横坐标上单击鼠标右键，选择“设置坐标轴格式→坐标轴选项”，“最小值、最大值、主要刻度单位、次要刻度单位”均选“固定”，值分别为 3、6.5、0.5、0.25。在“纵坐标轴交叉”框内，设定“坐标轴值”为 3。

按同样方法修改设置纵坐标的刻度。

④两次慢速单击横坐标轴标题，将标题改为“观察值”；两次慢速单击纵坐标轴标题，将标题改为“正态分位数”。

按以上 4 步进行修改，可获得正态 Q-Q 图（如图 15-6 右侧所示)。

P-P 图与 Q-Q 图还可用来检验其他的分布数据。这里只列举了检验正态分布的例子。

15.3 生存曲线

生存资料的特点就是有“删失（sensor)”，引起删失的原因通常有：观察对象仍然存活、死于与研究疾病或处理无关的因素、搬迁和其他各种原因退出研究等。为了充分利用删失信息，需要采用生存分析（survival analysis）方法。

例 15-3：某研究人员将 15 名某恶性肿瘤病人随机分为两个治疗组，第一组 8 人采

用实验新药治疗，第二组 7 人采用常规药治疗。然后随访两种药治疗肿瘤后病人的生存情况。生存时间及删失情况如表 15-1 所示，请绘制生存曲线图。

表 15-1　两种药治疗肿瘤后病人的生存时间

	生存时间（月）							
第一组	1.3	2	3	5	12^+	14^+	14^+	16^+
第二组	1	1.3	1.5	2	6	7	8	

注：表中数字后的“+”表示该个体删失。

	A	B	C	D	E	F	G	H	I	J
1	Kaplan-Meier法计算生存率及其标准误									
2										
3		生存		期内	期初	生存				生存率
4	编号	时间	删失	死亡数	病人数	概率	生存率	$\frac{d_i}{n_i(n_i-d_i)}$	$\sum_{t_i\le t}\frac{d_i}{n_i(n_i-d_i)}$	标准误
5	i	t_i	c_i	d_i	n_i	$p_i=1-d_i/n_i$	$\hat{S}(t)$			$SE[\hat{S}(t)]$
6	0	0	-	0	8	1.0000	1.0000	0.0000	0.0000	0.0000
7	1	1.3	0	1	8	0.8750	0.8750	0.0179	0.0179	0.1169
8	2	2	0	1	7	0.8571	0.7500	0.0238	0.0417	0.1531
9	3	3	0	1	6	0.8333	0.6250	0.0333	0.0750	0.1712
10	4	5	0	1	5	0.8000	0.5000	0.0500	0.1250	0.1768
11	5	12	1	0	4	1.0000	0.5000	0.0000	0.1250	0.1768
12	6	14	1	0	3	1.0000	0.5000	0.0000	0.1250	0.1768
13	7	14	1	0	2	1.0000	0.5000	0.0000	0.1250	0.1768
14	8	16	1	0	1	1.0000	0.5000	0.0000	0.1250	0.1768

图 15-7　生存率及其标准差的计算

15.3.1　Kaplan-Meier 法计算生存率及其标准误

Kaplan-Meier 法，也称为乘积极限法（product limit method）。如果令每个时间点 t_i 对应的期内死亡数为 d_i，期初病人数 n_i，生存率为 $\hat{S}(t)$，则生存率对应的标准误为：

$$SE[\hat{S}(t)]=\hat{S}(t)\sqrt{\sum_{t_i\le t}\frac{d_i}{n_i(n_i-d_i)}} \tag{15-1}$$

令时点 t_i 对应的生存概率为 $p_i=1-d_i/n_i$，则时点 t_i 对应的生存率为：

$$\hat{S}(t)=p_1p_2\cdots p_i \tag{15-2}$$

将生存时间 t_i 按从小到大顺序依次输入到单元格区域 B7:B106 中；时点 t_i 对应个体是否删失 c_i 输入到单元格区域 C7:C106 中，如果删失 c_i=1，否则 c_i=0（注意：如果某时点有 n 例删失，则 c_i=n）；每一个时点对应的期内死亡数为 d_i 输入到单元格区域 D7:D106；期初病例数在单元格 E6 中输入。这样程序自动得到生存率及其标准误。

对于例 15-3，将表 15-1 的数据按图 15-7 将已知数据分别输入到光盘文件“15-3 生存曲线.xlsx”的“生存率 1”、“生存率 2”工作表中，便由 Excel 自动获得生存率（见第 G 列）和它对应的标准误（见第 J 列）。

温馨提示：光盘文件“15-3 生存曲线.xlsx”的“生存率 1”、“生存率 2”工作表功能完全相同。用户只需修改工作表中单元格区域 B7:D106 以及单元格 E6 中的相关数据，便可获得所需要的生存率及其标准误。该程序最大可处理 100 个时点的生存时间数据。如果数据输入区原已有数据存在，最好先清除旧数据再输入新数据。

15.3.2 绘制生存曲线

将两组的生存时间与生存率复制、粘贴到“生存曲线”工作表中。选取“生存率 1”工作表中的单元格区域 B6:B14，按住 Ctrl 键，再选取单元格区域 G6:G14，复制、粘贴到“生存曲线”工作表的单元格 A3 处；选取“生存率 2”工作表中的单元格区域 B6:B13，按住 Ctrl 键，再选取单元格区域 G6:G13，复制、粘贴到“生存曲线”工作表的单元格 D3 处。

为了便于绘制阶梯形的生存曲线，需要对数据作一些处理。即添加一些阶梯线连接坐标点，如图 15-8 所示，灰色背景单元格的数据为新添数据。添加数据有如下规律：

① 生存时间为 0、生存率为 1 是曲线的起点；

② 在此坐标点后，如果生存率不同，则每隔一行插入一行数据；

③ 插入行的生存时间与下一行相同，生存率与上一行相同；

④ 删失数据所在行（如第一组的 11～14 行），以及最后一个坐标点位置无须添加新数据。

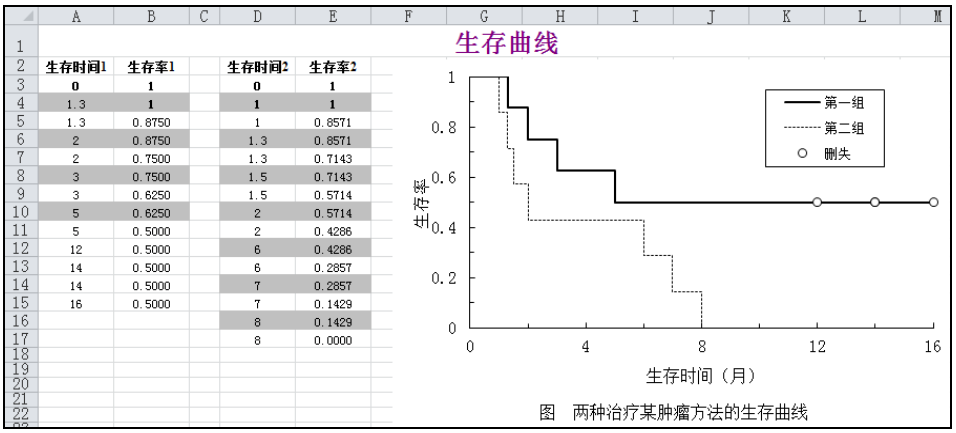


图 15-8 生存曲线及其绘制该图的有关数据

绘制图 15-8 可按以下步骤操作。

① 选择“插入→图表→*XY* 散点图”的“无数据点折线散点图”（第 5 个子图表类型）。

② 在弹出的图表框内单击鼠标右键选择“选择数据”。弹出“选择数据源”对话框，

删除所有系列，单击“添加”按钮。在“名称系列”右侧键入“第一组”，在“*X* 轴系列值”右侧用鼠标选取单元格区域 A3:A15，在“*Y* 轴系列值”右侧用鼠标选取单元格区域 B3:B15，单击“确定”按钮；再单击“添加”按钮，在“名称系列”右侧键入“第二组”，在“*X* 轴系列值”右侧用鼠标选取单元格区域 D3:D17，在“*Y* 轴系列值”右侧用鼠标选取单元格区域 E3:E17，单击“确定”按钮；再单击“添加”按钮，在“名称系列”右侧键入“删失”，在“*X* 轴系列值”右侧用鼠标选取单元格区域 A12:A15，在“*Y* 轴系列值”右侧用鼠标选取单元格区域 B12:B15，单击“确定”按钮。

③选中功能区“图表工具→布局→标签”后，单击“图表标题”，键入“图两种治疗某肿瘤方法的生存曲线”，再单击“坐标轴标题”，选择“主要纵/横坐标轴标题”后，在坐标轴下出现的标题框内分别填入“生存时间”和“生存率”。

④在横坐标上单击鼠标右键，选择“设置坐标轴格式→坐标轴选项”，主要刻度线类型与次要刻度线类型均选中“内部”，“最小值、最大值、主要刻度单位、次要刻度单位”均选“固定”，值分别为 0、16、4、2，单击“关闭”。

⑤在纵坐标上单击鼠标右键，选择“设置坐标轴格式→坐标轴选项”，主要刻度线类型与次要刻度线类型均选中“内部”，“最小值、最大值、主要刻度单位、次要刻度单位”均选“固定”，值分别为 0、1、0.2、0.1，单击“关闭”按钮。

⑥在“删失”数据系列标志上单击鼠标右键，数据标记类型选为内置，类型选择圆圈，大小改为 6 磅，圆圈颜色可以在“标记线颜色”中更改。单击“关闭”按钮。

⑦在“第一组”数据系列标志上单击鼠标右键，选择“设置数据系列格式→线条颜色→实线”，并选择适当宽度。在“第二组”数据系列标志上单击鼠标右键，选择“设置数据系列格式→线条颜色→实线”，并在“线型→短划线类型”中选择虚线（第三条）。单击“关闭”按钮。

⑧适当调整图例的形状及位置，调整标题位置，调整绘图区的大小，调整整个图表区域的字体大小等，可获得如图 15-8 右侧所示的生存曲线。

15.4 重复测量数据的曲线图

对同一研究对象（如人、动物、仪器等）的同一观测指标，在不同时间点的多次观测结果称为重复测量数据（repeated measures data），以此数据可反映某观测指标在不同时间点的变化情况。

例 15-4：为了比较新旧两种剂型的代谢情况，某研究者观测了 16 例某病患者服药后 0、4、8、12 小时的血药浓度（μmol/L）（如表 15-2 所示），试对该数据绘制新旧两种剂型的各个体的血药浓度-时间曲线。

表 15-2　4 个时点的某药新旧剂型血药浓度（μmol/L）

编号	旧剂型				编号	新剂型			
	0 小时	4 小时	8 小时	12 小时		0 小时	4 小时	8 小时	12 小时
1	90.53	142.12	65.54	73.28	8	70.53	97.38	112.12	58.50
2	88.43	163.17	48.95	71.77	9	68.43	95.27	133.17	56.90
3	100.01	144.75	86.06	80.01	10	57.37	78.43	83.16	48.34
4	46.32	126.33	48.95	39.54	11	105.80	120.54	136.33	84.03
5	73.69	138.96	70.02	60.89	12	80.01	104.75	114.75	65.61
6	105.27	126.33	75.01	83.66	13	56.32	75.27	96.33	47.52
7	86.32	121.06	78.95	70.24	14	53.69	110.02	138.96	45.44
					15	85.27	110.01	126.33	69.47
					16	66.32	115.27	129.06	55.29

解：绘制曲线时，以时间为横轴，血药浓度为纵轴；一条曲线代表一个个体，用不同的线条形状区分新旧药物两种剂型。其图形绘制步骤与前两节（15.2 和 15.3）一致，只不过其“*X* 轴系列值”右侧用鼠标选取单元格区域为 C3:F3，“*Y* 轴系列值”右侧用鼠标选取单元格区域也分别为横向的 C4：F4、C5：F5……C16:F16。

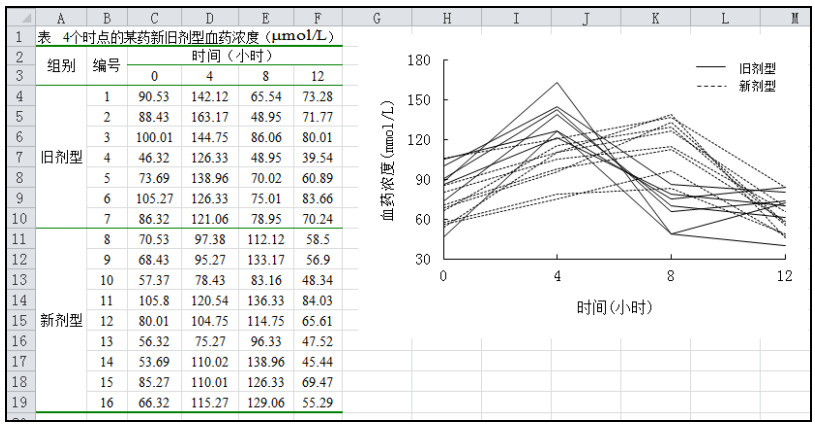


图 15-9　4 个时点的某药新旧剂型血药浓度（μmol/L）

结果分析：如图 15-9 所示，在时间点 0、12 处新旧剂型两组差异较小，而在时间点 4、8 处新旧剂型两组差异较大。时间因素对新旧剂型两组之间的差异起了重要作用。

15.5　Meta 分析统计量的置信区间图

Meta 分析（meta analysis）可以说是一种定量的观察研究方法，用于合并研究内容相

同的多个独立研究的信息，所以也叫“研究中的研究（studies of studies）”。通常情况下假定效应值（effect size）近似服从正态分布。

15.5.1 Meta 分析的计算公式

最常见的 Meta 分析效应值（effect size）有三种：一是连续资料效应值，如均数之差；二是离散资料效应值，如优势比（Odds Ratio，OR）的对数、率比（相对危险度，Relative Risk，RR）的对数、率差（危险差，Risk Difference，RD），其中危险差（RD）的倒数称为需要治疗病人数（Number Needed to Treat，NNT），是临床决策中很有用的一项指标，即需要治疗多少个病人才能预防 1 次不利结局出现；三是 Pearson 相关系数 r，除了 Pearson 相关系数 r 外，可将假设检验统计量 t、F 和 χ^2 值以及 Spearman 相关系数 r_s 通过类似 $r_i=\sqrt{t_i^2/(t_i^2+df_i)}$ 公式进行变换（如表 15-3 所示），其中 t_i 为第 i 研究的 t 检验统计量，df_i 为检验统计量相应的自由度。

表 15-3　转换成相关系数的公式与方法

统计量	转换公式	说明
t	$r=\sqrt{t^2/\left(t^2+df\right)}$，$df=n_1+n_2-2$	适用于成组 t 检验，也适用于配对 t 检验（$df=n-1$）
F	$r=\sqrt{F/\left(F+df_{组内}\right)}$	适用于单因素方差分析
F	$r=\sqrt{F_a df_a/\left(F_a df_a+F_b df_b+F_{ab} df_{ab}+df_{误差}\right)}$	适用于双因素（a、b）析因分析
χ^2	$r=\sqrt{\chi^2/n}$	适用于自由度为 1 的卡方检验，n 为总例数

假定共有 k 个研究，连续资料研究 i（$i=1,2,\cdots,k$）的试验组例数、均数、标准差分别记为 n_{1i}、$\bar{x}_{1i}$、s_{1i}，对照组例数、均数、标准差分别记为 n_{2i}、$\bar{x}_{2i}$、s_{2i}；离散资料的试验（暴露）组阳性、试验（暴露）组阴性、对照组阳性、对照组阴性分别记为 a_i、b_i、c_i、d_i，试验（暴露）组、对照组阳性率分别记为 p_{1i}、p_{0i}，试验（暴露）组、对照组总例数分别记为 n_{1i}、n_{0i}；相关系数 r 类资料每个研究的总例数记为 n_i，相关系数 r_i 经正态变换后记为 Z_{ri}。Meta 分析时，除 r_i 需要正态变换外，OR 和 RR 值也需要取对数进行正态变换。取对数后的 OR、RR 值分别记为 LOR 和 LRR。

主要分析结果有固定效应（fixed effect）的合并效应值（pooled effect size）及其标准误与置信区间、随机效应（random effect）的合并效应值及其标准误与置信区间。一般只需选择其中之一进行结果解释，选择的标准是齐性检验（homogeneity test）Q 值，Q 值服从自由度为 k-1（k 为研究数）的 χ^2 分布，如果 Q 值较小，相应 P 值较大，则说明各研究资料间变异较小，可认为 Meta 分析研究资料为“齐性”，采用固定效应结果；如果 Q 值较大，相应 P 值较小，则说明各研究资料间变异较大，可认为 Meta 分析研究资料“不齐”，应采用随机效应结果。

进行 Meta 分析利用的具体计算公式如表 15-4a 和表 15-4b 所示。

表 15-4a Meta分析的计算公式

	连续资料效应值	相关系数
效应值（effect size）	$\bar{X}_i = \bar{X}_{1i} - \bar{X}_{2i}$	$z_{r_i} = 0.5\ln\frac{1+r_i}{1-r_i} = \tanh^{-1} r_i$ （r 值变换参见表 15-3）
标准误	$SE_{\bar{X}_i} = \sqrt{\frac{(n_{1i}-1)S_{1i}^2 + (n_{2i}-1)S_{2i}^2}{(n_{1i}+n_{2i}-2)}\left(\frac{1}{n_{1i}}+\frac{1}{n_{2i}}\right)}$	$SE_{z_{r_i}} = \sqrt{\frac{1}{n_i-3}}$
固定效应权重	$w_i = 1/SE_{\bar{X}_i}^2$	$w_i = 1\Big/SE_{z_{r_i}}^2 = n_i - 3$
固定效应合并效应值	$\bar{X} = \frac{\Sigma w_i \bar{X}_i}{\Sigma w_i}$	$z_r = \frac{\Sigma w_i z_{r_i}}{\Sigma w_i}$
固定效应合并标准误	$SE_{\bar{X}} = \sqrt{1/\Sigma w_i}$	$SE_{Z_r} = \sqrt{1/\Sigma w_i}$
齐性检验	$Q = \Sigma w_i\left(\bar{X}_i - \bar{X}\right)^2$，$\nu = k-1$，$i = 1,2,\cdots,k$	$Q = \Sigma w_i\left(z_{r_i} - z_r\right)^2$
随机效应权重	$w_i^* = 1/(1/w_i + SE_\tau^2)$，其中 $SE_\tau^2 = \max(0, \frac{Q-k+1}{\Sigma w_i - \Sigma w_i^2/\Sigma w_i})$	
随机效应合并效应值	$\bar{X}_R = \frac{\Sigma w_i^* \bar{X}_i}{\Sigma w_i^*}$	$z_{rR} = \frac{\Sigma w_i^* z_{ri}}{\Sigma w_i^*}$
随机效应合并标准误	$SE_{\bar{X}_R} = \sqrt{1/\Sigma w_i^*}$	$SE_{z_{rR}} = \sqrt{1/\Sigma w_i^*}$

表 15-4b Meta分析的计算公式

	离散资料（包括 OR、RR、RD）
效应值 （effect size）	$\mathrm{LOR}_i = \ln\frac{p_{1i}/(1-p_{1i})}{p_{0i}/(1-p_{0i})} = \ln\frac{a_i d_i}{b_i c_i}$，$\mathrm{LRR}_i = \ln\frac{p_{1i}}{p_{0i}} = \ln\frac{a_i}{n_{1i}}\Big/\frac{c_i}{n_{0i}}$ $\mathrm{RD}_i = p_{1i} - p_{0i} = \frac{a_i}{n_{1i}} - \frac{c_i}{n_{0i}}$（$n_{1i} = a_i + b_i$、$n_{0i} = c_i + d_i$）
标准误	$SE_{\mathrm{LOR}_i} = \sqrt{\frac{1}{n_{1i}p_{1i}(1-p_{1i})} + \frac{1}{n_{0i}p_{0i}(1-p_{0i})}} = \sqrt{\frac{1}{a_i}+\frac{1}{b_i}+\frac{1}{c_i}+\frac{1}{d_i}}$ $SE_{\mathrm{LRR}_i} = \sqrt{\frac{1-p_{1i}}{n_{1i}p_{1i}} + \frac{1-p_{0i}}{n_{0i}p_{0i}}} = \sqrt{\frac{1}{a_i} - \frac{1}{n_{1i}} + \frac{1}{c_i} - \frac{1}{n_{0i}}}$ $SE_{\mathrm{RD}_i} = \sqrt{\frac{p_{1i}(1-p_{1i})}{n_{1i}} + \frac{p_{0i}(1-p_{0i})}{n_{0i}}} = \sqrt{\frac{a_i b_i}{n_{1i}^3} + \frac{c_i d_i}{n_{0i}^3}}$
固定效应 权重	$w_i = 1/SE_{\mathrm{LOR}_i}^2$，$w_i = 1/SE_{\mathrm{LRR}_i}^2$，$w_i = 1/SE_{\mathrm{RD}_i}^2$
固定效应 合并效应值	$\mathrm{LOR} = \frac{\Sigma w_i \mathrm{LOR}_i}{\Sigma w_i}$，$\mathrm{LRR} = \frac{\Sigma w_i \mathrm{LRR}_i}{\Sigma w_i}$，$\mathrm{RD} = \frac{\Sigma w_i \mathrm{RD}_i}{\Sigma w_i}$
固定效应 合并标准误	$SE_{\mathrm{LOR}} = \sqrt{1/\Sigma w_i}$，$SE_{\mathrm{LRR}} = \sqrt{1/\Sigma w_i}$，$SE_{\mathrm{RD}} = \sqrt{1/\Sigma w_i}$
齐性检验	$Q = \Sigma w_i\left(\mathrm{LOR}_i - \mathrm{LOR}\right)^2$，$Q = \Sigma w_i\left(\mathrm{LRR}_i - \mathrm{LRR}\right)^2$ $Q = \Sigma w_i\left(\mathrm{RD}_i - \mathrm{RD}\right)^2$

续表

	离散资料（包括 OR、RR、RD）
随机效应 权重	$w_i^* = 1/(1/w_i + SE_\tau^2)$，其中 $SE_\tau^2 = \max\left(0, \dfrac{Q-k+1}{\Sigma w_i - \Sigma w_i^2/\Sigma w_i}\right)$
随机效应 合并效应值	$\mathrm{LOR}_R = \dfrac{\Sigma w_i^* \mathrm{LOR}_i}{\Sigma w_i^*}$，$\mathrm{LRR}_R = \dfrac{\Sigma w_i^* \mathrm{LRR}_i}{\Sigma w_i^*}$，$\mathrm{RD}_R = \dfrac{\Sigma w_i^* \mathrm{RD}_i}{\Sigma w_i^*}$
随机效应 合并标准误	$SE_{\mathrm{LOR}_R} = \sqrt{1/\Sigma w_i^*}$，$SE_{\mathrm{LRR}_R} = \sqrt{1/\Sigma w_i^*}$，$SE_{\mathrm{RD}_R} = \sqrt{1/\Sigma w_i^*}$

齐性检验的 Q 值实际上为各研究效应值与合并效应值间的加权离差平方和，即

$$Q = \sum \text{各研究权重}\left(\text{各研究效应值} - \text{合并效应值}\right)^2$$

反映了各研究间变异（inter-study variation）情况，如果 Q 值较小，则可忽略研究间变异，只考虑各研究内变异（inner-study variation）（如 $SE_{\bar{X}}$、SE_{Z_r}、SE_{LOR}、SE_{LRR}、SE_{RD}），采用固定效应模型结果；如果 Q 值较大，则不可忽略研究间变异 SE_τ，采用随机效应模型结果。随机效应模型除考虑研究内变异外，还考虑研究间变异，因此其结果比固定效应模型更保守，相应置信区间更宽。

Meta 分析的统计分析步骤如下：

① 齐性检验

H_0：各研究来自同一总体，即各研究效应值的总体水平相同。

H_1：各研究来自不同总体，即各研究效应值的总体水平不全相同。

② 合并效应值的点估计及区间估计

不同类型效应值的点估计及区间估计的计算方法有所差异，见各实例的叙述。一般假定效应值服从正态分布，其 95%置信区间为：效应值 ±1.96标准误 。

③ 合并效应值的检验

H_0：总体合并效应值=0（$\bar{X}$、z_r、RD）或=1（LOR、LRR）

H_1：总体合并效应值≠0（$\bar{X}$、z_r、RD）或≠1（LOR、LRR）

$$z = \text{合并效应值}/\text{标准误}$$

如果 $z \geqslant 1.96$，$P \leqslant 0.05$，则拒绝 H_0，说明合并效应值有统计学意义，否则无统计学意义。

④ 绘制置信区间图和漏斗图

置信区间图可直观反映每一研究效应值，以及合并效应值在一定置信度下的总体置信区间范围。漏斗图（funnel plot）的横坐标为各研究的效应值，纵坐标为各研究的样本含量（或与样本大小有密切关系的标准误），如果漏斗图呈现对称倒置漏斗状，则无发表偏倚，否则认为有发表偏倚。

Meta 分析时常假定：

① k 个研究之间相互独立（注意同一课题组的研究常常很难保持独立）；

②选择研究时应具体地、明确地规定所抽取研究的纳入标准和排除标准，以确保所选研究是同类研究；

③所抽取的研究是采用随机方法抽取的，无出版偏倚或抽样偏倚；

④k 个研究的设计正确，资料整理、分析、解释合理。

15.5.2　连续资料的 Meta 分析

例 15-5：某研究者通过 MEDLINE 查得 9 份主题相同的研究，每份研究均有试验组例数、均数、标准差，对照组例数、均数、标准差数据（如图 15-10 所示），试根据该数据进行 Meta 分析，并绘制相应的置信区间图、漏斗图。

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	连续型资料的Meta分析													
2	请输入数据！									分析结果				
3	研究	试验组				对照组				研究个数k		9		
4	编号	例数n_{1i}	均数$\bar{X}_{1i}$	标准差S_{1i}^2		例数n_{2i}	均数$\bar{X}_{2i}$	标准差S_{2i}^2		齐性检验Q值	5.402	自由度	8	
5	1	134	5.96	4.24		113	6.82	4.72		P值	0.71385194	不拒绝H0，齐性成立		
6	2	175	4.74	4.64		151	5.07	5.38						
7	3	137	2.04	2.59		140	2.51	3.22				均数之差	标准误	Z值
8	4	184	2.7	2.32		179	3.2	2.46		固定效应		-0.2834971	0.092273009	3.0723725
9	5	174	6.09	4.86		169	5.81	5.14		随机效应		-0.2834971	0.092273009	3.0723725
10	6	754	4.72	5.33		736	4.76	5.29						
11	7	209	10.1	8.1		209	10.9	7.9				下限	上限	
12	8	1151	2.82	3.05		1122	3.01	3.32		固定效应合并差值的95%CI		-0.4643522	-0.102641958	
13	9	679	3.88	4.85		673	4.37	5.37		随机效应合并差值的95%CI		-0.4643522	-0.102641958	
14														
15											最终结果			
16												均数之差	标准误	
17											固定效应	-0.2834971	0.092273009	
18												下限	上限	
19											固定效应合并差值的95%CI	-0.4643522	-0.102641958	
20											Z值	3.0723725	P值	0.002124
21													合并效应值有统计学意义	

图 15-10　均数之差的 Meta 分析数据与结果

解：光盘文件“15-5 连续型 Meta 分析.xlsx”有“Meta 分析”、“置信区间图”、“漏斗图”3 个工作表，其中“Meta 分析”工作表分为数据输入区（单元格区域 A2:H104）、Meta 分析计算结果区（单元格区域 J2:N19）（如图 15-10 所示）及其中间结果计算区（第 P 至 AB 列）（如图 15-11 所示）。

1. Meta 分析

中间结果计算区又分成固定效应模型、齐性检验、随机效应模型计算区，本例在图 15-11 的第 U 列获得了固定效应模型获得的合并效应值（均数之差）为–0.2835，相应标准误为 0.0923，z 值为 3.0724>1.96，表明合并效应值有统计学意义。齐性检验 Q 值为 5.4022，在

自由度=8 的卡方分布中 P 等于 0.7139（即“=CHIDIST(5.4022,8)”），说明各研究间齐同，无须采用随机效应模型结果。从图 15-11 的右侧可见，随机效应模型获得的合并效应值、标准误及其 z 值与固定效应模型结果几乎等同，表明研究间变异 SE_τ 十分小，可以忽略。这些主要结果已经总结在图 15-10 的右侧，特别是第 15 行至第 21 行的“最终结果”。

	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB
1											$SE_\tau^2=\max(0,\frac{Q-k+1}{\sum w_i-\sum w_i^2/\sum w_i})$		
2	研究	固定效应						齐性检验		随机效应			
3	编号	均数之差	标准误	权重		合并效应值		Q		权重			合并效应值
4	i	$\bar{X}_i=\bar{X}_{i1}-\bar{X}_{i2}$	$SE_{\bar{X}_i}$	$w_i=1/SE_{\bar{X}_i}^2$	$w_i\bar{X}_i$	$\bar{X}=\Sigma w_i\bar{X}_i/\Sigma w_i$		$Q=\Sigma w_i(\bar{X}_i-\bar{X})^2$		w_i^2	$w_i^*=1/(1/w_i+SE_\tau^2)$	$w_i^*\bar{X}_i$	$\bar{X}_R=\Sigma w_i^*\bar{X}_i/\Sigma w_i^*$
5	合计			117.4494	-33.296546	-0.283497056		5.40217922		3825.86	117.4493542	-33.29654612	-0.283497056
6	1	-0.86	0.5703737	3.073838	-2.6435008			1.021607448		9.44848	3.073838114	-2.643500778	
7	2	-0.33	0.5549378	3.247218	-1.0715819	标准误		0.007022186		10.5444	3.247217991	-1.071581937	标准误
8	3	-0.47	0.3515663	8.090691	-3.8026249	$SE_{\bar{X}}=\sqrt{1/\Sigma w_i}$		0.28142133		65.4593	8.090691231	-3.802624878	$SE_{\bar{X}_R}=\sqrt{1/\Sigma w_i^*}$
9	4	-0.5	0.2509146	15.88357	-7.941787	0.092273009		0.744519099		252.288	15.88357397	-7.941786986	0.092273009
10	5	0.28	0.5399966	3.429398	0.96023156			1.088933216		11.7608	3.429398417	0.960231557	
11	6	-0.04	0.2751604	13.20773	-0.5283091	Z值		0.783096972		174.444	13.20772799	-0.52830912	Z值
12	7	-0.8	0.7826467	1.632557	-1.3060459	3.072372502		0.435525979		2.66524	1.632557413	-1.30604593	3.072372502
13	8	-0.19	0.1336688	55.96801	-10.633922			0.489255522		3132.42	55.96800998	-10.6339219	
14	9	-0.49	0.2782469	12.91634	-6.3290062			0.550797466		166.832	12.91633908	-6.329006151	

图 15-11　均数之差的 Meta 分析计算公式与中间结果

2. 95%置信区间图

光盘文件“15-5 连续型 Meta 分析.xlsx”的“置信区间图”工作表界面如图 15-12 所示。第 A 列是研究编号，为了便于绘图这里将“合计”以“0”编号，从第 2 研究编号起，以后的编号会根据第 B 列有无数据自动生成；第 B 列合计的均数之差公式为“=Meta 分析!L17”（见单元格 B3），是根据齐性检验 P 值是否大于 0.05 而从固定效应与随机效应中选择的合并效应值，从第 1 研究编号起的均数之差等于“Meta 分析”工作表的第 Q 列；单元格 C3 公式为“=1.96*Meta 分析!M17”，“M17”是根据齐性检验结果从“Meta 分析”工作表中选取的标准误，从第 1 研究编号起的值等于 1.96 乘以“Meta 分析”工作表的第 R 列；第 D、E 两列的 95%置信区间下、上限根据第 B、C 两列获得。权重百分率是各研究在计算加权平均值（即合并效应值）时所用权重的百分构成比。一般情况下，样本含量越大，标准误越小，权重就越大，即所得权重百分率越大。

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	总体均数之差的95%置信区间												
2	研究编号	均数之差	1.96标准误	95%CI下限	95%CI上限	权重百分率(%)							
3	0	-0.283497	0.180855098	-0.464352154	-0.10264196								
4	1	-0.86	1.117932484	-1.977932484	0.257932484	2.617161							
5	2	-0.33	1.087678012	-1.417678012	0.757678012	2.764781							
6	3	-0.47	0.689069866	-1.159069866	0.219069866	6.888664							
7	4	-0.5	0.491792562	-0.991792562	-0.00820744	13.52376							
8	5	0.28	1.058393344	-0.778393344	1.338393344	2.919896							
9	6	-0.04	0.539314396	-0.579314396	0.499314396	11.24547							
10	7	-0.8	1.533987567	-2.333987567	0.733987567	1.39001							
11	8	-0.19	0.261990859	-0.451990859	0.071990859	47.65289							
12	9	-0.49	0.545363857	-1.035363857	0.055363857	10.99737							
13													

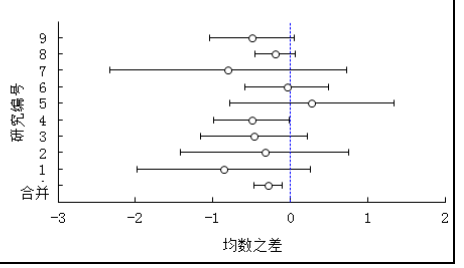


图 15-12　均数之差的置信区间图

Meta 分析均数之差的 95%总体置信区间图形可按下列步骤进行绘制。

①选择“插入→图表→其他图表→所有图表类型→*XY* 散点图”。

②在弹出的图表框内单击鼠标右键选择“选择数据”。弹出“选择数据源”对话框，删除所有系列，单击“添加”。在“*X* 轴系列值”右侧用鼠标选取单元格区域 B3:B12，在“*Y* 轴系列值”右侧用鼠标选取单元格区域 A3:A12，如图 15-13 所示，单击“确定”按钮。

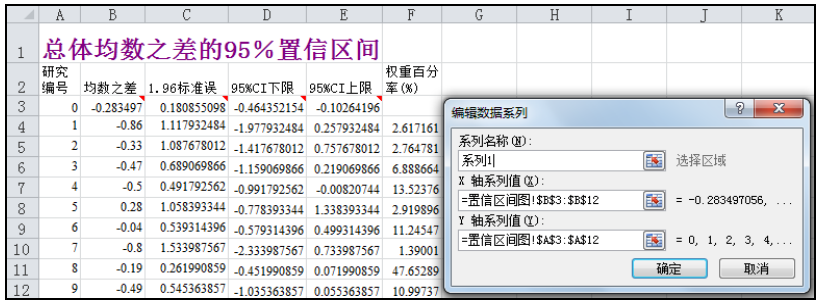


图 15-13 置信区间图的源数据添加

③删除图例、数值轴主要网格线、绘图区背景颜色。

④在纵坐标上单击鼠标右键，选择“设置坐标轴格式→坐标轴选项”，主要刻度线类型选“内部”，“最小值、最大值、主要刻度单位”均选“固定”，值分别为-1、9.99（研究数+0.99，以便上方不至于太紧凑）、1；选择“数字”，在分类下拉列表中选择数值，小数位数设置为 0 位，负数选择红色“1234.10”（第 3 个），“坐标轴值”键入-1。单击“关闭”按钮。在横坐标上单击鼠标右键设置 *X* 轴，选择“坐标轴选项→坐标轴值”，键入–3，单击“关闭”按钮。

⑤在数据系列（散点）标志上单击鼠标右键，选择“设置数据系列格式→数据标记选项”数据标记类型选为内置，类型选择圆圈，大小改为 5 磅，圆圈颜色可以在“标记线颜色”中更改。单击“关闭”按钮。

⑥在功能区栏中选择“图表工具→布局→分析→误差线→标准误差误差线”，弹出的“设置水平误差线”中“误差量”自定义均选取单元格区域 C3:C12（1.96 标准误）。

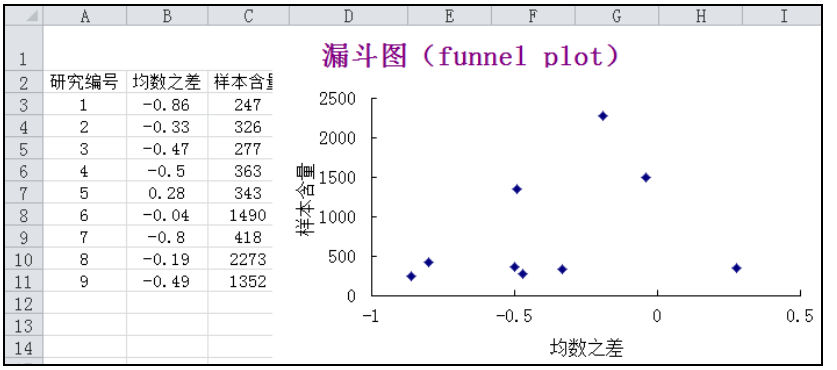
⑦在“图表工具→布局→标签→坐标轴标题”中，在主要横/纵坐标轴标题内分别填入“均数之差”、“研究编号”。

⑧为了使图中显示 *X*=0 的纵线，可单击鼠标右键选择“选择数据”。弹出“选择数据源”对话框，单击“添加”按钮。在“*X* 轴系列值”右侧空白处输入“0, 0”，在“*Y* 轴系列值”右侧空白处输入“–1, 9.99”（9.99 为研究数+0.99），然后在这两点组成的直线处单击鼠标右键，修改其数据系列格式。

⑨为了使 *Y* 轴下部 0、1 处显示“合并”，可单击功能区“图表工具→布局→插入→文本框”进行设定。绘制好的完成图形如图 15-12 右侧所示。

3. 漏斗图

这里均数之差为“Meta 分析”工作表的第 Q 列，样本含量为“Meta 分析”工作表数据区的第 B 列与第 F 列数据之和（即 $n_{1i}+n_{2i}$）。漏斗图是以均数之差为横轴，样本含量为纵轴绘制的一种散点图，如图 15-14 所示。



研究编号	均数之差	样本含量
1	-0.86	247
2	-0.33	326
3	-0.47	277
4	-0.5	363
5	0.28	343
6	-0.04	1490
7	-0.8	418
8	-0.19	2273
9	-0.49	1352

图 15-14　均数之差的漏斗图

结果分析：由图 15-12 可见，尽管 9 个研究中有 8 个研究（除研究编号 4 外）的均数之差 95%置信区间包括 0，但合并效应值的 95%置信区间为（–0.4643,–0.1026），不包括 0，说明试验组与对照组间的差异具有统计学意义（z =3.0724，P=0.0011）。图 15-14 所显示的漏斗图不太对称（图中右下角的点相对较少），提示可能存在一定的发表偏倚。

温馨提示：该程序最大可输入 100 个研究的数据。只需在“Meta 分析”工作表的单元格区域 B5:D104（试验组）与 F5:H104（对照组）中相应位置输入已知的各研究例数、均数、标准差便可获得所有结果。但必须注意，如果研究数不同（即不为 9 个研究），则 95%置信区间图以及漏斗图的相应源数据做必要改变。如数据输入区原已有数据存在，最好先清除旧数据再输入新数据。

对于 95%置信区间图，需要修改“选择数据→编辑数据系列”的“系列 1”X 值单元格区域结尾单元格“B12”与 Y 值单元格区域结尾单元格“A12”中的数值，如 10 个研究，则该数字由 12 改变为 13 即可。修改系列 3 的 Y 值“9.99”，如 10 个研究，则 Y 值“9.99”改变为“10.99”。此外，还需修改误差线的数据，如 10 个研究，则修改“设置水平误差线”中“误差量”自定义选取单元格区域 C3:C13。

对于漏斗图，需要修改“选择数据→编辑数据系列”的“系列 1”X 值单元格区域结尾单元格“B11”与 Y 值单元格区域结尾单元格“C11”中的数值，如 10 个研究，则该数字由 11 改变为 12 即可。

15.5.3 离散资料的 Meta 分析

根据研究的设计类型或需要，离散资料又可分为优势比（Odds Ratio，OR）、相对危险度（Relative Risk，RR）和危险差（Risk Difference，RD）资料。

1. 优势比

例 15-6：Pap 试验是一种收集、制备和检查子宫颈脱落细胞的试验，因为该试验快速、价廉、无创伤，所以医生常用它作为早期宫颈癌的筛选。某研究者通过 MEDLINE 查得 13 份有关 Pap 试验的研究，每份研究的试验组例数、阳性率，对照组例数、阳性率分别如图 15-15 所示，试根据该数据进行 Meta 分析。

	A	B	C	D	E	F	G	H	I	J	K	L
1	优势比OR的Meta分析											
2	请输入数据！									分析结果		
3		试验(病例)组			对照组				研究个数k	13		
4	研究编号	例数n_{1i}	阳性率p_{1i}		例数n_{0i}	阳性率p_{0i}			齐性检验Q值	34.09667203	自由度	12
5	1	710	0.8014085		278	0.4244604			P值	0.000651264	拒绝H0，齐性不成立	
6	2	43	0.1395349		8	0.01						
7	3	194	0.371134		41	0.3170732				优势比		
8	4	34	0.2941176		91	0.043956			固定效应	4.150318964		
9	5	8	0.5		27	0.1111111			随机效应	3.768979614		
10	6	63	0.7460317		87	0.4137931						
11	7	77	0.4285714		21	0.1904762				下限	上限	
12	8	39	0.6923077		59	0.6440678			固定效应总OR的95%CI	3.407241697	5.05545219	
13	9	74	0.9054054		20	0.55			随机效应总OR的95%CI	2.526769486	5.62188494	
14	10	58	0.6896552		69	0.4202899						
15	11	58	0.637931		252	0.3690476				最终结果		
16	12	42	0.5238095		11	0.3636364				优势比		
17	13	153	0.2679739		282	0.0319149			随机效应	3.768979614		
18										下限	上限	
19									随机效应总OR的95%CI	2.526769486	5.62188494	
20									Z值	6.50357342	P值	7.843E-11
21											合并效应值有统计学意义	

图 15-15 优势比 Meta 分析的数据与分析结果

解：光盘文件“15-6 优势比 Meta 分析.xlsx”有“OR”、“置信区间图”、“漏斗图”3 个工作表，其中“OR”工作表又分为数据输入区、Meta 分析计算结果区及其中间结果计算区。优势比 Meta 分析需要的基本数据是试验组与对照组各自的例数与阳性率，其效应值为优势比的对数，所以优势比的 95%置信区间时计算公式与连续资料 Meta 分析不同，即先对服从正态分布的 LOR 求置信区间，然后对 LOR 置信区间的上、下限取指数，即总体 OR 的 95%置信区间为：

$$\exp(\mathrm{LOR}_i \pm 1.96SE_{\mathrm{LOR}_i}) \qquad (15\text{-}3)$$

效应值 LOR_i 及其标准误 SE_{LOR_i} 的计算也比较特殊（如表 15-4b 所示），如第 1 研究的 LOR_1 在单元格 O6 的计算公式为“=IF(N6<>"",LN((C5/(1−C5))/(F5/(1−F5))),"")”，将此公式向下拖放至单元格 O105 可获得其他研究的优势比对数值；第 1 研究的 SE_{LOR_1} 在单元格 P6

的计算公式为“=IF(N6<>"",(1/(B5*C5*(1–C5))+(1/(E5*F5*(1–F5))))^0.5,"")”，将此公式向下拖放至单元格 P105 可获得其他研究的优势比对数的标准误值。此后的其他计算公式与连续资料 Meta 分析相似。此外，因为 OR 不服从正态分布，只是 LOR 服从正态分布，对 LOR 进行 z 检验的公式不能用于 OR。LOR 值的固定效应模型、齐性检验、随机效应模型计算公式及计算结果如图 15-16 所示。

	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1													$SE_\tau^2=\max(0,\frac{Q-k+1}{\Sigma w_i-\Sigma w_i^2/\Sigma w_i})$
2	研究	固定效应						齐性检验		随机效应			
3	编号	优势比对数	标准误	权重		合并效应指标		Q		权重			合并效应指标
4	i	LOR_i	SE_{LOR_i}	$w_i=1/SE^2_{\mathrm{LOR}_i}$	$w_i\mathrm{LOR}_i$	$\mathrm{LOR}=\Sigma w_i\mathrm{LOR}_i/\Sigma w_i$		$Q=\Sigma w_i(\mathrm{LOR}_i-\mathrm{LOR})^2$		w_i^2	$w_i^*=1/(1/w_i+SE_\tau^2)$	$w_i^*\mathrm{LOR}_i$	$\mathrm{LOR}_R=\Sigma w_i^*\mathrm{LOR}_i/\Sigma w_i^*$
5	合计			98.704448	140.47471	1.42318519		34.09667203		2181.658299	24.02649117	31.8784519	1.326804305
6	1	1.699609735	0.15353906	42.419176	72.096044			3.241271635		1799.386459	3.204749557	5.44682354	
7	2	2.775961407	3.58049666	0.0780034	0.2165344	标准误		0.142746469		0.006084528	0.076286851	0.21176935	标准误
8	3	0.239900227	0.36704628	7.4226391	1.7806928	$SE_{\mathrm{LOR}}=\sqrt{1/\Sigma w_i}$		10.39290683		55.09557056	2.363029409	0.56689129	$SE_{\mathrm{LOR}_R}=\sqrt{1/\Sigma w_i^*}$
9	4	2.20414502	0.63494954	2.4803991	5.4671594	0.100654139		1.512791113		6.152379917	1.445873288	3.18691441	0.204011582
10	5	2.079441542	0.93541435	1.1428571	2.3765046			0.492197027		1.306122449	0.859503334	1.78728694	
11	6	1.425865574	0.36216325	7.6241468	10.871009	Z值		5.47753E-05		58.12761508	2.383081062	3.39795325	Z值
12	7	1.15923691	0.60154288	2.7635468	3.2036055	14.1393608		0.192532697		7.637190905	1.537712905	1.78257356	6.503573423
13	8	0.217866494	0.44080064	5.1465425	1.1212592			7.476861674		26.48689927	2.07138907	0.45128627	
14	9	2.058111775	0.59983558	2.7793008	5.7201118			1.120424465		7.724513215	1.542578235	3.17479843	
15	10	1.12009132	0.37421693	7.1409029	7.9984633			0.656005426		50.99249384	2.333717203	2.61397638	
16	11	1.102700184	0.30280038	10.906544	12.026648			1.120218127		118.9527066	2.630535799	2.90069231	
17	12	0.654925968	0.69879302	2.0478723	1.3412048			1.208699785		4.193781123	1.287375226	0.84313547	
18	13	2.407320413	0.38482844	6.7525168	16.255472			6.539962013		45.59648327	2.290659233	5.51435073	

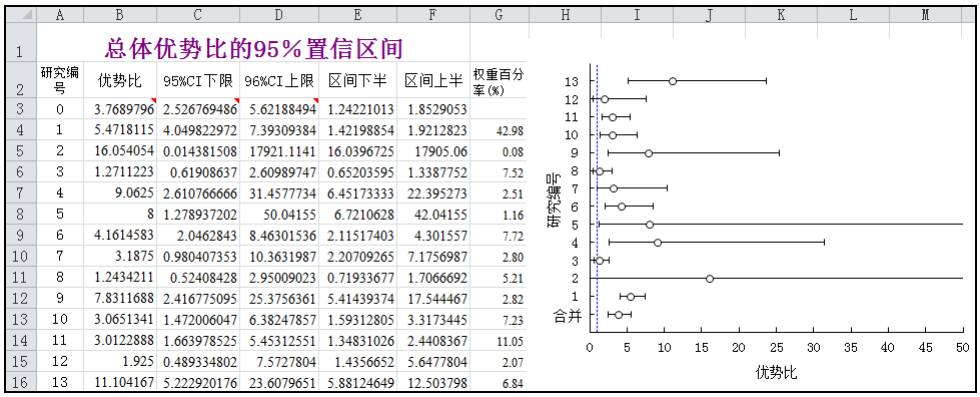
图 15-16　优势比对数的 Meta 分析计算公式与中间结果

光盘文件“15-6 优势比 Meta 分析.xlsx”的“置信区间图”工作表界面如图 15-17 所示。第 A 列是研究编号，为了便于绘图，这里将“合计”以“0”编号，从第 2 研究编号起，以后的编号会根据第 B 列有无数据自动生成；第 B 列合计的优势比公式为“=OR!J17”（见单元格 B3），是根据齐性检验 P 值是否大于 0.05 而从固定效应与随机效应中选择的合并效应值，从第 1 研究编号起的优势比等于“OR”工作表的第 O 列之指数，如单元格 B4 的计算公式为“=IF(OR!N6<>"",EXP(OR!O6),"")”；单元格 C3、D3 也是根据齐性检验 P 值是否大于 0.05 而从固定效应与随机效应中选择的合并效应值 95%置信区间下、上限值，单元格区域 C4:C103 与 D4:D103 是各研究效应值的 95%置信区间下、上限值（根据公式（15-3）计算）。第 E、F 列是根据第 B、C、D 列相应数据获得的 95%置信区间下一半区间宽度、上一半区间宽度，即 95%置信区间下、上误差线端点值。权重百分率与前面类似。尽管第 2 研究的置信区间较宽，但权重百分率仅占 0.08%，第 1 个研究的权重百分比最大。

Meta 分析优势比的 95%总体置信区间图是以第 B 列优势比为横坐标，第 A 列研究编号为纵坐标绘制的散点图，自定义误差线–、+值分别为第 E、F 列。因为试验组与对照组的效应完全相同时的优势比为 1，所以垂直参照线的 X=1（注意：连续资料 Meta 分析的垂直参照线为 0）漏斗图的绘制与连续资料 Meta 分析相似。

结果分析：本例获得齐性检验结论为“不齐”（Q=34.0967，df=12，P=0.000651），采用随机效应模型结果，即该例优势比合并值为 3.7690，相应的 95%置信区间为

（2.5268,5.6219），不包括 0，表明 Pap 试验对早期宫颈癌具有一定筛选效果。



总体优势比的95%置信区间

研究编号	优势比	95%CI下限	96%CI上限	区间下半	区间上半	权重百分率(%)
0	3.7689796	2.526769486	5.62188494	1.24221013	1.8529053	
1	5.4718115	4.049822972	7.39309384	1.42198854	1.9212823	42.98
2	16.054054	0.014381508	17921.1141	16.0396725	17905.06	0.08
3	1.2711223	0.61908637	2.60989747	0.65203595	1.3387752	7.52
4	9.0625	2.610766666	31.4577734	6.45173333	22.395273	2.51
5	8	1.278937202	50.04155	6.7210628	42.04155	1.16
6	4.1614583	2.0462843	8.46301536	2.11517403	4.301557	7.72
7	3.1875	0.980407353	10.3631987	2.20709265	7.1756987	2.80
8	1.2434211	0.52408428	2.95009023	0.71933677	1.7066692	5.21
9	7.8311688	2.416775095	25.3756361	5.41439374	17.544467	2.82
10	3.0651341	1.472006047	6.38247857	1.59312805	3.3173445	7.23
11	3.0122888	1.663978525	5.45312551	1.34831026	2.4408367	11.05
12	1.925	0.489334802	7.5727804	1.4356652	5.6477804	2.07
13	11.104167	5.222920176	23.6079651	5.88124649	12.503798	6.84

图 15-17　优势比的 Meta 分析置信区间图

温馨提示：该程序最大可输入 100 个研究的数据。只需在“OR”工作表的单元格区域 B5:C104（试验组）与 E5:F104（对照组）中相应位置输入已知的各研究例数、均数、标准差便可获得所有结果。如数据输入区原已有数据存在，最好先清除旧数据再输入新数据。但必须注意，如果研究数不同（即不为 13 个研究），则 95%置信区间图以及漏斗图的相应源数据做必要改变，请参照连续资料 Meta 分析的温馨提示。

2. 相对危险度

相对危险度 Meta 分析方法与上面所述优势比十分相似，也是先将相对危险度 RR 转换为服从正态分布的相对危险度之对数 LRR，然后对 LRR 进行类似连续资料 Meta 分析的方法进行分析。

光盘文件“15-7 相对危险度 Meta 分析.xlsx”有“RR”、“置信区间图”、“漏斗图”3 个工作表，其中“RR”工作表又分为数据输入区、Meta 分析计算结果区及其中间结果计算区。效应值 LRR_i 及其标准误 SE_{LRR_i} 的计算也比较特殊（如表 15-4b 所示），如第 1 研究的 LRR_1 在单元格 O6 的计算公式为“=IF(N6<>"",LN(C5/F5),"")”，将此公式向下拖放至单元格 O105 可获得其他研究的相对危险度对数值；第 1 研究的 SE_{LRR_1} 在单元格 P6 的计算公式为“=IF(N6<>"",((1−C5)/(B5*C5)+(1−F5)/(E5*F5))^0.5,"")”，将此公式向下拖放至单元格 P105 可获得其他研究的相对危险度对数的标准误值。此后的其他计算公式、95%置信区间图的绘制及修改、漏斗图的绘制及修改与优势比 Meta 分析极其相似。

采用例 15-6 优势比数据，进行相对危险度 Meta 分析，由分析结果可见，优势比合并效应值（3.7690）与相对危险度合并效应值（1.8878）相差较大，且置信区间有较大差别（如图 15-18 所示），这是因为阳性率较高的缘故。如果阳性率很低，优势比结果与相对危险度结果会很接近。

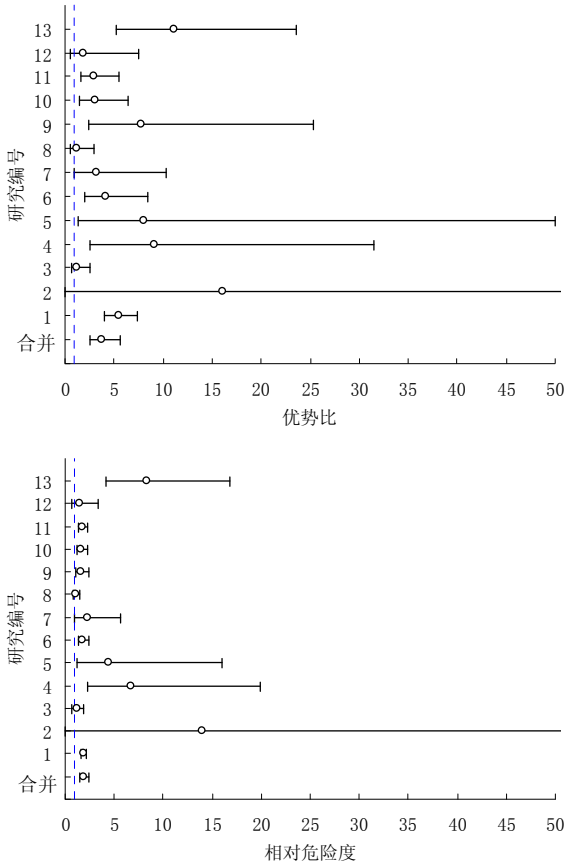


图 15-18　同一资料优势比（左）与相对危险度（右）的 95%置信区间

3. 危险差

危险差资料尽管也属于离散资料，但与上述优势比、相对危险度不同，危险差本身服从正态分布，不用变换。

光盘文件“15-8 危险差 Meta 分析.xlsx”有“RD”、“置信区间图”、“漏斗图”3 个工作表，其中“RD”工作表又分为数据输入区、Meta 分析计算结果区及其中间结果计算区。效应值 RD_i 及其标准误 SE_{RD_i} 的计算也比较特殊（如表 15-4b 所示），如第 1 研究的 RD_1 在单元格 O6 的计算公式为“=IF(N6<>"",C5–F5,"")”，将此公式向下拖放至单元格 O105 可获得其他研究的危险差；第 1 研究的 SE_{RD_1} 在单元格 P6 的计算公式为“=IF(N6<>"",(C5*(1–C5)/B5+F5*(1–F5)/E5)^0.5,"")”，将此公式向下拖放至单元格 P105 可获得其他研究的相对危险度对数的标准误值。此后的其他计算公式、95%置信区间图的绘制及修改、漏斗图的

绘制及修改与连续资料 Meta 分析极其相似。

对以往 6 个研究进行危险差 Meta 分析，获得如图 15-19 所示的分析结果，该数据齐性检验 P=0.1492>0.05，表明研究间齐同，采用固定效应结果。合并效应值危险差= –0.2890，95%置信区间为（–0.3469,–0.2311），结果有统计学意义。

危险差RD的Meta分析

请输入数据！					分析结果			
试验（病例）组			对照组		研究个数k	6		
研究编号	例数n_{1i}	阳性率p_{1i}	例数n_{0i}	阳性率p_{0i}	齐性检验Q值	8.130	自由度	5
1	69	0.45	72	0.68	P值	0.1492317	不拒绝H0，齐性成立	
2	224	0.13	102	0.54				
3	45	0.62	14	0.86		危险差	标准误	Z值
4	48	0.65	13	0.85	固定效应	-0.289032	0.02954977	9.78118547
5	206	0.57	71	0.86	随机效应	-0.275478	0.03994172	6.89699216
6	61	0.67	58	0.84				
						下限	上限	
					固定效应总RD的95%CI	-0.346949	-0.2311142	
					随机效应总RD的95%CI	-0.353763	-0.1971919	
					最终结果			
						危险差	标准误	
					固定效应	-0.289032	0.02954977	
						下限	上限	
					固定效应总RD的95%CI	-0.346949	-0.2311142	
					Z值	9.781185	P值	0
					合并效应值有统计学意义			

图 15-19　危险差的 Meta 分析数据及其分析结果

15.5.4　相关系数资料的 Meta 分析

可通过反双曲正切函数 $z_{r_i}=\tanh^{-1}r_i$ 变换，使相关系数 r_i 变成服从正态分布的效应值 z_{r_i}，效应值对应的标准误为 $SE_{Z_{r_i}}=\sqrt{1/(n_i-3)}$，其他计算，诸如固定效应、随机效应合并效应值、齐性检验的计算类似其他 Meta 分析（如图 15-20 所示）。

研究编号	固定效应					齐性检验	随机效应			
	正态变换	标准误	权重		合并效应值		权重			合并效应值
i	$Z_{r_i}=\tanh^{-1}r_i$	$SE_{Z_{r_i}}$	$w_i=1/SE_{Z_{r_i}}^2$	$w_iZ_{r_i}$	$Z_r=\Sigma w_iZ_{r_i}/\Sigma w_i$	$Q=\Sigma w_i(Z_{r_i}-Z_r)^2$	w_i^2	$w_i^*=1/(1/w_i+SE_\tau^2)$	$w_i^*Z_{ri}$	$Z_{rR}=\Sigma w_i^*Z_{r_i}/\Sigma w_i^*$
合计			999	403.321768	0.403725494	27.38226421	126453	254.3769666	103.8633778	0.408304962
1	0.562729769	0.0883883	128	72.0294105		3.236142038	16384	32.07658125	18.05044717	
2	0.522984278	0.0890871	126	65.896019	标准误	1.792054849	15876	31.94949413	16.70908311	标准误
3	0.309519604	0.0811107	152	47.0469798	$SE_{Z_r}=\sqrt{1/\Sigma w_i}$	1.348961942	23104	33.3980816	10.337361	$SE_{Z_R}=\sqrt{1/\Sigma w_i^*}$
4	0.213171347	0.0920575	118	25.1542189	0.0316386	4.284684193	13924	31.40953251	6.69561234	0.062699071
5	0.693147181	0.096225	108	74.8598955		9.046610584	11664	30.65401748	21.24774579	
6	0.497311288	0.0928477	116	57.6881094	Z值	1.015962894	13456	31.26604161	15.54895541	Z值
7	0.223656109	0.0957826	109	24.3785159	12.76053604	3.53432318	11881	30.73404823	6.873857642	6.512137327
8	0.255412812	0.0839181	142	36.2686193		3.123524526	20164	32.88916982	8.400315345	

$SE_\tau^2=\max(0,\frac{Q-k+1}{\Sigma w_i-\Sigma w_i^2/\Sigma w_i})$

图 15-20　相关系数资料的 Meta 分析计算公式与中间结果

由 8 个以往研究的各研究例数与相关系数数据（如图 15-21 所示），得到合并相关系数（通过双曲正切函数 $r=\tanh z_r$ 变换获得）为 0.3870（齐性检验 P<0.05,取随机效应结果），95%置信区间为（0.2779，0.4863），经检验差异有统计学意义，即置信度=95%时总体相关

系数不包括无相关性的 0 值。8 个以往研究相关系数及合并相关系数的 95%置信区间如图 15-22 所示。

	A	B	C	D	E	F	G	H	I
1	相关系数资料的Meta分析								
2		请输入数据！					分析结果		
3	研究编号	例数n_i	相关系数r_i			研究个数k	8		
4	1	131	0.51			齐性检验Q值	27.382	自由度	7
5	2	129	0.48			P值	0.00028439	拒绝H0，齐性不成立	
6	3	155	0.3						
7	4	121	0.21				相关系数		
8	5	111	0.6			固定效应	0.38313212		
9	6	119	0.46			随机效应	0.3870325		
10	7	112	0.22						
11	8	145	0.25				下限	上限	
12						固定效应总ρ的95%CI	0.27367801	0.482789847	
13						随机效应总ρ的95%CI	0.27790914	0.486294133	
14									
15							最终结果		
16							相关系数		
17						随机效应	0.3870325		
18							下限	上限	
19						随机效应总r的95%CI	0.27790914	0.486294133	
20						Z值	6.5121373	P值	7.41E-11
21								合并效应值有统计学意义	

图 15-21　相关系数资料的 Meta 分析与结果

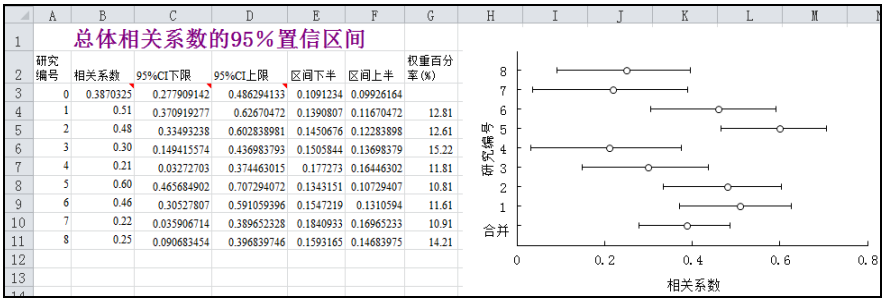


图 15-22　总体相关系数的 95%置信区间

相关系数的置信区间是先按正态分布获得 z_{r_i} 的 95%置信区间，然后对其进行双曲正切函数变换获得，即

$$\tanh(z_{r_i} \pm 1.96SE_{z_{r_i}}) \qquad (15\text{-}4)$$

注意：因为相关系数等于 0 的垂直参考线正好与纵坐标轴重合，所以在图 15-22 中无法看到这条直线。

15.5.5　*P* 值的 Meta 分析

有时，以往研究已知信息为具体 *P* 值，或具有可转换成具体 *P* 值的有关信息，此种情

况下可采用 P 值的 Meta 分析。具有检验统计量 t、F、χ^2 及其自由度时，可分别通过“TDIST（t 值，自由度，1）”、“FDIST（F 值，分子自由度，分母自由度）/2”、“CHIDIST（卡方值，自由度）/2”进行单侧 P 值转换。P 值的 Meta 分析有两种方法：一是通过公式

$$\chi^2 = -2\sum_{i=1}^{k}\ln(P_i), \qquad \nu = 2k \tag{15-5}$$

进行卡方检验，其中 P_i 为各研究单侧 P 值，ν 为自由度，k 为研究数；二是通过公式

$$z = \sum_{i=1}^{k}\frac{z(P_i)}{\sqrt{k}} \tag{15-6}$$

进行正态 z 检验，其中 $z(P_i)$ 是由各研究单侧 P 值获得正态离差值的。

24 个以往研究的单侧 P 值职图 15-23 所示的单元格区域 B6:B29，由卡方检验和 z 检验分别获得检验 P 值为 1.9942E-11 和 2.2022E-14，表明 24 个研究的合并效应有统计学意义。

	A	B	C	D	E	F	G	H	I	J	K
1	P值Meta分析										
2					卡方检验				Z检验		
3	研究编号	单侧P_i值	研究数		卡方检验值	P值	卡方临界值		Z检验值	P值	Z临界值
4	i	P_i	k		$\chi^2=-2\sum_{i=1}^{k}\ln(P_i)$		$\chi^2_{0.05,2k}$		$Z=\sum_{i=1}^{k}Z(P_i)/\sqrt{k}$		$Z_{0.05}$
5	0		24		143.3314384	1.99417E-11	65.1707689		-7.548458434	2.2022E-14	-1.644853627
6	1	0.1406			-1.9618363				-1.077627555		
7	2	0.0412			-3.189317023				-1.736927323		
8	3	0.0333			-3.402197882				-1.834363863		
9	4	0.3476			-1.056702886				-0.391808113		
10	5	0.1298			-2.041760475				-1.127337058		
11	6	0.0001			-9.210340372				-3.719016485		
12	7	0.0256			-3.665162927				-1.949799678		
13	8	0.1067			-2.237734121				-1.244270566		
14	9	0.0923			-2.382711137				-1.326723997		
15	10	0.0321			-3.438899249				-1.850788404		
16	11	0.0899			-2.409057338				-1.341371087		
17	12	0.2134			-1.54458694				-0.794679431		
18	13	0.2202			-1.513219055				-0.771517928		
19	14	0.1824			-1.701553201				-0.906256697		
20	15	0.1467			-1.919365594				-1.050692156		
21	16	0.0476			-3.044922518				-1.668583253		
22	17	0.0029			-5.843044542				-2.758879035		
23	18	0.0193			-3.947650183				-2.068425813		
24	19	0.1206			-2.115275995				-1.171992668		
25	20	0.0234			-3.755019257				-1.98810571		
26	21	0.0731			-2.615926912				-1.453085559		
27	22	0.0655			-2.725705136				-1.510170202		

图 15-23 P 值的 Meta 分析

第 16 章

诊断试验评价

随着科学技术的发展，循证医学理念越来越受到关注，诊断试验评价的使用越来越广泛。本章主要介绍如何使用 Excel 来分析、处理诊断试验评价数据。首先介绍常用统计指标，再介绍 ROC 分析原理与曲线绘制，然后介绍诊断试验的 Meta 分析、诊断试验的样本含量估计，最后介绍诊断一致性的估计方法。

16.1 常用统计指标

诊断试验（diagnostic test）评价以金标准（gold standard）划分病例与对照（分别记为D_+与D_-），诊断测量结果写成阳性与阴性（分别记为T_+与T_-），金标准与诊断测量间相互独立。金标准是指当前医学界公认的诊断某一疾病最可靠的诊断方法。常用的金标准有病理学诊断、外科手术发现、影像学诊断，以及公认的综合临床诊断标准、长期临床随访结果等。

例 16-1：采用 ECG（心电图）对具有急性持久胸痛的 700 名患者进行诊断，经证实有 520 例出现心肌梗塞，其余 180 例没有出现心肌梗塞，结果如表 16-1 所示。试计算 ECG 诊断试验的几个常用评价指标。

表 16-1 ECG诊断试验的结果

ECG 诊断结果	心肌梗塞		合 计
	出 现	不 出 现	
阳性	415（*TP*）	10（*FP*）	425
阴性	105（*FN*）	170（*TN*）	275
合计	520	180	700（*N*）

表 16-2 诊断资料 2×2 四格表

诊断结果（*T*）	金标准（*D*）		合 计
	病例（D_+）	非病例(D_-)	
阳性（T_+）	*TP*（真阳性）	*FP*（假阳性）	*TP*+*FP*
阴性（T_-）	*FN*（假阴性）	*TN*（真阴性）	*FN*+*TN*
合计	*TP*+*FN*	*FP*+*TN*	*N*

表 16-1 数据可用表 16-2 的符号表示，四格表数据的 4 个可能结果中，2 个是正确的，即病例被诊断为阳性（真阳性）和非病例被诊断为阴性（真阴性）；2 个是错误的，即病例被诊断为阴性（假阴性，或漏诊）和非病例被诊断为阳性（假阳性，或误诊）。诊断试验评价常用指标及其标准误、95%置信区间的计算公式如表 16-3 所示，常用指标间的相互关系如图 16-1 所示。

表 16-3 常用诊断试验指标的计算公式

名 称	指 标	标 准 误	95%CI	备 注
粗一致率	$Acc=\frac{TP+TN}{N}\times 100\%$	$SE_{Acc}=\sqrt{\frac{Acc\times(1-Acc)}{N}}$	$Acc\pm 1.96SE_{Acc}$	依赖于患病率

续表

名　称	指　标	标 准 误	95%CI	备　注
诊断一致性 Kappa 值	$\kappa=\frac{p_O-p_E}{1-p_E}$（$p_O=Acc$，观察正确率，$p_E$ 期望正确率）	$SE_\kappa=\frac{1}{(1-p_E)\sqrt{N}}\times\sqrt{p_E+p_E^2-\frac{\sum R_iC_i(R_i+C_i)}{N^3}}$ $P_E=\sum R_iC_i/N^2$	$\kappa\pm1.96SE_\kappa$（左侧 R_i、C_i 分别为第 i 行合计、第 i 列合计，N 为总例数）	一般只在金标准缺乏情况下使用，取值范围为（-1,1）
灵敏度	$Sen=P(T_+\mid D_+)=TPR$ $=TP/(TP+FN)$	$SE_{Sen}=\sqrt{\frac{Sen\times(1-Sen)}{TP+FN}}$	$Sen\pm1.96SE_{Sen}$	与病例组有关
特异度	$Spe=P(T_-\mid D_-)$ $=TN/(FP+TN)$	$SE_{Spe}=\sqrt{\frac{Spe\times(1-Spe)}{FP+TN}}$	$Spe\pm1.96SE_{Spe}$	与对照组有关
Youden 指数	$J=Sen+Spe-1$ $=TPR-FPR$	$SE_J=\sqrt{SE_{Sen}^2+SE_{Spe}^2}$	$J\pm1.96SE_J$	取值范围为（-1,1），越接近 1 越好
阳性似然比	$LR_+=TPR/FPR$ $=Sen/(1-Spe)$	$SE_{LR+}=e^{\sqrt{\frac{1-Sen}{TP}+\frac{Spe}{FP}}}$	$\frac{Sen}{1-Spe}e^{\pm1.96\sqrt{\frac{1-Sen}{TP}+\frac{Spe}{FP}}}$	取值范围为（0,∞），值越大越好
阴性似然比	$LR_-=(1-Sen)/Spe$ $=(1-TPR)/(1-FPR)$	$SE_{LR-}=e^{\sqrt{\frac{Sen}{FN}+\frac{1-Spe}{TN}}}$	$\frac{1-Sen}{Spe}e^{\pm1.96\sqrt{\frac{Sen}{FN}+\frac{1-Spe}{TN}}}$	取值范围同上，值越小越好
优势比	$OR=\frac{Sen/(1-Sen)}{(1-Spe)/Spe}$ $=\frac{TP\times TN}{FP\times FN}=\frac{LR_+}{LR_-}$	$SE_{OR}=e^{\sqrt{\frac{1}{TP}+\frac{1}{FP}+\frac{1}{TN}+\frac{1}{FN}}}$	$OR\times e^{\pm1.96\sqrt{\frac{1}{TP}+\frac{1}{FP}+\frac{1}{TN}+\frac{1}{FN}}}$	与非病例相比，病例的阳性优势大小
阳性预测价值	$PV_+=P(D_+\mid T_+)$ $=TP/(TP+FP)$ $=1/\left(1+\frac{(1-Spe)(1-P_0)}{SenP_0}\right)$	$SE_{PV+}=\sqrt{\frac{PV_+\times(1-PV_+)}{TP+FP}}$（左侧 P_0 为患病率，或称先验概率）	$PV_+\pm1.96SE_{PV+}$	取值范围为（0,1），值越接近 1 越好
阴性预测价值	$PV_-=P(D_-\mid T_-)$ $=TN/(TN+FN)$ $=1/\left(1+\frac{(1-Sen)P_0}{Spe(1-P_0)}\right)$	$SE_{PV-}=\sqrt{\frac{PV_-\times(1-PV_-)}{TN+FN}}$（左侧 P_0 为患病率，或称先验概率）	$PV_-\pm1.96SE_{PV-}$	同上

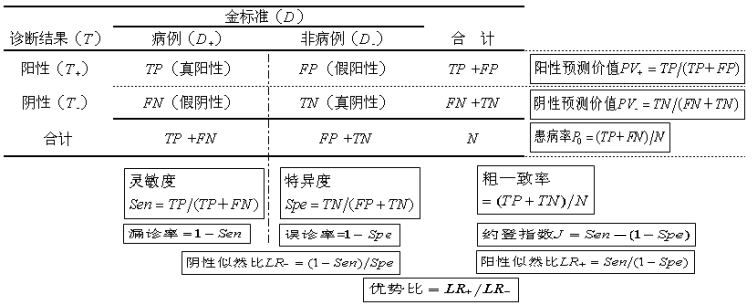


图 16-1　常用指标间的相互关系

1. 一致性指标

表 16-3 共列出了 10 个常用诊断试验指标，其中前两个指标为一致性指标。粗一致率（crude accuracy）是病例正确诊断为阳性以及非病例正确诊断为阴性的两者之和占总例数的比例。诊断一致性 Kappa 值是观察一致率（observed probability of agreement，$P_O = Acc$）与期望一致率（expected probability of agreement，P_E）之差占 1 与期望一致率之差的比例，即 $k=(P_O-P_E)/(1-P_E)$，其中期望一致率 $P_E=\sum R_iC_i/N^2$，R_i、C_i 分别为第 i 行合计、第 i 列合计，N 为总例数。这两个一致性指标是非常粗略的诊断准确度评价指标，一般情况下较少使用。原因是一致性的好与坏，并不能完全说明试验诊断性能的好与坏。与金标准的一致性好，当然说明试验诊断好；但如果一致性较差，并不能完全说明某诊断试验诊断性能差。如果被错误诊断的患者在两个试验中差异明显，即使诊断准确度较高，也不会有较好的一致性。这两个指标除受患病率、诊断临界值的影响外，还受漏诊率与误诊率的影响。Kappa 值的理论取值在–1～1 范围内，Kappa 值的绝对值为 0～0.4 时说明一致性程度不理想；Kappa 值的绝对值大于等于 0.75 时说明具有较好的一致性。

2. 灵敏度与特异度

灵敏度（sensitivity）与特异度（specificity）分别是真阳性率与真阴性率，不受患病率影响，其取值范围均在（0, 1）之间，其值越接近于 1，说明其诊断准确度越好，两者之间的关系如图 16-2 所示。病例组曲线下诊断临界值右侧为灵敏度，左侧为漏诊率，漏诊率=1-灵敏度；非病例组曲线下诊断临界值左侧为特异度，右侧为误诊率，误诊率=1-特异度。

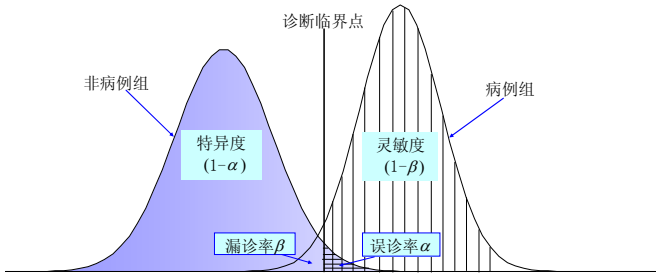


图 16-2 灵敏度、特异度漏诊率与误诊率

3. 合并诊断指标

当比较两个诊断试验时，单独使用灵敏度或特异度，可能出现一个诊断试验的灵敏度高、特异度低，而另一个诊断试验的灵敏度低、特异度高，无法判断哪一个诊断试验更好。由此，需要有将灵敏度和特异度结合的诊断试验评价指标，这些指标有 Youden 指数、阳性似然比、阴性似然比、优势比等。Youden 指数（Youden's Index）为真阳性率与假阳性率之差，阳性似然比（positive likelihood ratio，LR_+）为真阳性率与假阳性率之比，阴性似然比

（negative likelihood ratio，LR_-）为假阴性率与真阴性率之比，优势比（odds ratio）是病例阳性优势与非病例阳性优势之比。似然比大小及其相应含义如表 16-4 所示。

表 16-4　似然比大小及其对应的意义

LR_+	LR_-	意　义
>10	<0.1	引起较大改变
5～10	0.1～0.2	引起中等改变
2～5	0.2～0.5	引起较小改变
<2	>0.5	引起微弱改变

优势（Odds）为两互斥概率之比，先验优势（Pre-test Odds）=先验概率/(1–先验概率)，后验优势（Post-test Odds）=先验优势×似然比，其中先验概率$=(TP+FN)/(TP+FN+FP+TN)$，后验概率=后验优势/(1+后验优势)。

4. 预测价值指标

当诊断试验结果是阳性时，受试者实际为病例的概率就是阳性预测价值（positive predictive value），说明受试者真正有病的概率有多大；当诊断试验结果是阴性时，受试者实际为非病例的概率就是阴性预测价值（negative predictive value）。这两个指标受患病率的影响较大，当灵敏度与特异度为常数时，患病率增加，阳性预测价值增加，阴性预测价值减少。

根据表 16-3 中公式编制的光盘文件“16-1 诊断试验评价指标.xlsx”，可计算上述诊断试验评价指标。对于例 16-1 数据，其文件部分界面如图 16-3 所示，第 3～8 行为基本数据输入区，单元格区域 C6:D7 为四格表实际频数，可直接键入或滑动滚动条改变格子值，其右侧表格用于格子输入内容提示。此外，单元格 G10 可自定义置信区间的置信度，单元格 D27 可自定义先验概率（如人群患病率）。

结果分析：心电图诊断心肌梗塞的灵敏度、特异度分别为 79.81%（76.36%, 83.26%）（指标值后括号内为 95%CI，以下同）、94.44%（91.10%, 97.79%），Youden 指数为 74.25%（69.45%, 79.06%），阳性似然比为 14.37（7.85, 26.28），阴性似然比为 0.21（0.18, 0.25），均说明诊断效果比较好。优势比为 67.1905（34.2877, 131.6669），表明出现心肌梗塞者的 ECG 诊断结果阳性优势是不出现心肌梗塞的 67 倍。以上指标是固有指标，不会随患病率的变化而变化。

阳性预测价值为 97.65%（96.21%, 99.09%），即 97.65%诊断阳性者出现心肌梗塞，说明诊断阳性确诊患病的能力强；阴性预测价值为 61.82%（56.08%, 67.56%），说明 61.82%诊断阴性者没有出现心肌梗塞，但还有 38.18%诊断阴性者中出现心肌梗塞，提示该诊断试验阴性时排除疾病的能力不强。预测价值随患病率变化而变化，如果人群患病率只有万分

之五（P_0=0.0005=0.05%），则阳性预测价值改变为 0.71%（0, 1.51%），即 100 个诊断阳性者只有不到 1 人出现心肌梗塞，说明如果将心电图用于人群心肌梗塞诊断，阳性确诊患病的能力相当差。



请输入四格表实际频数！

试验		金标准 D+	金标准 D-	合计
	T+	415	10	425
	T-	105	170	275
合计		520	180	700

诊断资料2×2四格表

诊断结果（T）	金标准（D） 病例（D+）	非病例（D-）	合计
阳性（T+）	TP（真阳性）	FP（假阳性）	TP+FP
阴性（T-）	FN（假阴性）	TN（真阴性）	FN+TN
合计	TP+FN	FP+TN	N

	诊断指标	标准误	95.00% 置信区间 下限	上限	P_0	P_E
一致率	0.835714286	0.0140049	0.8082652	0.863163		
Kappa值	0.633257403	0.0360201	0.5626593	0.703856	0.835714	0.552041
灵敏度	0.798076923	0.0176041	0.7635735	0.83258		
特异度	0.944444444	0.0170732	0.9109815	0.977907		
漏诊率	0.201923077	0.0176041	0.1674197	0.236426		
误诊率	0.055555556	0.0170732	0.0220926	0.089018		
Youden指数	0.742521368	0.0245234	0.6944563	0.790586		
阳性似然比	14.36538462	1.360849	7.8533438	26.27725		
阴性似然比	0.213800905	1.0931207	0.1795649	0.254564		
优势比（odds ratio）	67.19047619	1.4095113	34.287727	131.6669		
阳性预测值	0.976470588	0.0073526	0.9620598	0.990881		
阴性预测值	0.618181818	0.0292968	0.5607612	0.675602		
先验概率（如人群患病率）	0.05%					
阳性预测值	0.007135011	0.0040827	0	0.015137		
阴性预测值	0.999893058	0.0006236	0.9986709	1		

图 16-3 诊断试验评价指标的数据输入与分析结果界面

在缺乏金标准的情况下，可参考诊断一致性结果，例 16-1 的诊断一致率为 83.57%（80.83%,86.82%），Kappa 值为 0.6333（0.5627, 0.7039），说明诊断试验准确度尚可。

温馨提示：只要在单元格区域 C6:D7 中直接键入或滑动滚动条改变四格表实际频数值，便可获得所有诊断指标的计算结果。注意在单元格 G10 中可自定义置信区间的置信度，在单元格 D27 中可自定义先验概率（如人群患病率）。

16.2 ROC 曲线原理及其绘制方法

由图 16-2 可知，如果将诊断临界值左右移动，同一项检测方法，将有不同的灵敏度与（1-特异度）。为了更全面地评价检测方法的诊断价值，有必要考虑各种可能的诊断临界值。ROC 曲线是接收者工作特征（Receiver Operating Characteristic，ROC）曲线或相对工作特征（Relative Operating Characteristic）曲线的缩写。ROC 分析于 20 世纪 50 年代起源于统计决策理论，后来应用于雷达信号接收能力的评价。自从 20 世纪 80 代起，该方法广泛应用于医学诊断试验性能的评价。

以（1-特异度）即假阳性率（False Positive Rate，FPR）为横轴，灵敏度即真阳性率（True Positive Rate，TPR）为纵轴可绘制 ROC 曲线。ROC 曲线一定通过（0,0）和（1,1）两个坐标点，依次连接各点可得到经验 ROC 曲线（empirical ROC curve），中间每一个坐标点的斜率即阳性似然比（如果检测值越大越有可能阳性，这是通常情况）。图 16-2 中诊断临界值从左向右移动，灵敏度、（1-特异度）逐渐减少，ROC 曲线从右上角逐渐滑向左下角，阳性似然比逐渐增大。完全无价值的诊断有 TPR=FPR，是一条从原点到右上角的对角线，这条线称为机会线（chance line）；ROC 曲线一般位于机会线的上方，离机会线越远，说明诊断准确度越高；最好的诊断试验在图中表现为 ROC 曲线从原点垂直上升至左上角，然后水平到达右上角。

ROC 曲线下面积（记为 A_z）可反映诊断试验的准确度大小。这一指标取值范围在 0.5 至 1 之间，完全无价值的诊断 A_z=0.5，完全理想的诊断 A_z=1。一般认为 A_z 在 0.5～0.7 之间表示诊断准确度较低；在 0.7～0.9 之间表示诊断准确度为中等；0.9 以上表示诊断准确度较高。

在光盘中有两个文件“16-2 ROC 曲线.xlsx”、“16-3 小值阳性 ROC.xlsx”，可以计算 ROC 曲线下面积及其标准误，此外还可自动绘制 ROC 曲线，前者适用于检测值越大越有可能是阳性的一般情况，后者适用于检测值越小越有可能是阳性的特殊情况。曲线下面积及其面积的计算见公式（16-1）与公式（16-2）。

$$A_z=\frac{1}{n_a n_n}\sum_{j=1}^{n_n}\sum_{i=1}^{n_a}\psi(X_{a_i},X_{n_j}),\quad \psi(X_{a_i},X_{n_j})=\begin{cases}1 & X_{a_i}>X_{n_j}\\ 0.5 & X_{a_i}=X_{n_j}\\ 0 & X_{a_i}<X_{n_j}\end{cases} \tag{16-1}$$

$$SE_{A_z}=\sqrt{\frac{A_z(1-A_z)+(n_a-1)(Q_1-A_z^2)+(n_n-1)(Q_2-A_z^2)}{n_a n_n}} \tag{16-2}$$

病例组与非病例组的样本含量分别为 n_a、n_n，X_{a_i}、X_{n_j} 分别为病例组第 i 个体与非病例组第 j 个体的检测值。SE_{A_z} 为面积 A_z 的标准误，Q_1 是两个随机选择的病例组观察值比一个随机选择的非病例组观察值都将有更大可能分类为病例的概率，Q_2 是一个随机选择的病例组观察值比两个随机选择的非病例组观察值将有更大可能分类为病例的概率。由公式

$$z=\left(A_z-0.5\right)/SE_{A_z} \tag{16-3}$$

来检验 ROC 曲线下面积是否与完全随机情况下获得的 A_z=0.5 有统计学差异，A_z 的 95%置信区间为：

$$A_z\pm 1.96SE_{A_z} \tag{16-4}$$

下面利用 3 个实例分别说明 3 种不同的 ROC 分析数据。

16.2.1 有序分类资料 ROC 分析

例 16-2：有 109 份 CT 影像，其中有 51 份采用金标准确诊为异常，58 份确诊为正

常。某放射科医生在不知道“金标准”的前提下，根据这些 CT 影像的异常程度将 109 份影像判归为“很可能正常、可能正常、异常可疑、可能异常、很可能异常”5 类，依次记为 1、2、3、4、5，诊断分类结果如表 16-5 所示。试计算所有可能的灵敏度与（1-特异度）值，绘制 ROC 曲线图，并计算 ROC 曲线下面积。

表 16-5 109 份CT影像分类

金 标 准	诊断分类					合 计
	1	2	3	4	5	
异常	3	2	2	11	33	51
正常	33	6	6	11	2	58

解：打开光盘文件“16-2 ROC 曲线.xlsx”，先清除单元格区域 C6:D10 中的内容，然后将“正常”组数据依次键入到单元格区域 C6:C10，将“异常”组数据依次键入到单元格区域 D6:D10，程序自动产生的结果如图 16-4 所示。其中第 E 至 I 列为中间计算结果，最终结果见第 K 列和第 L 列。

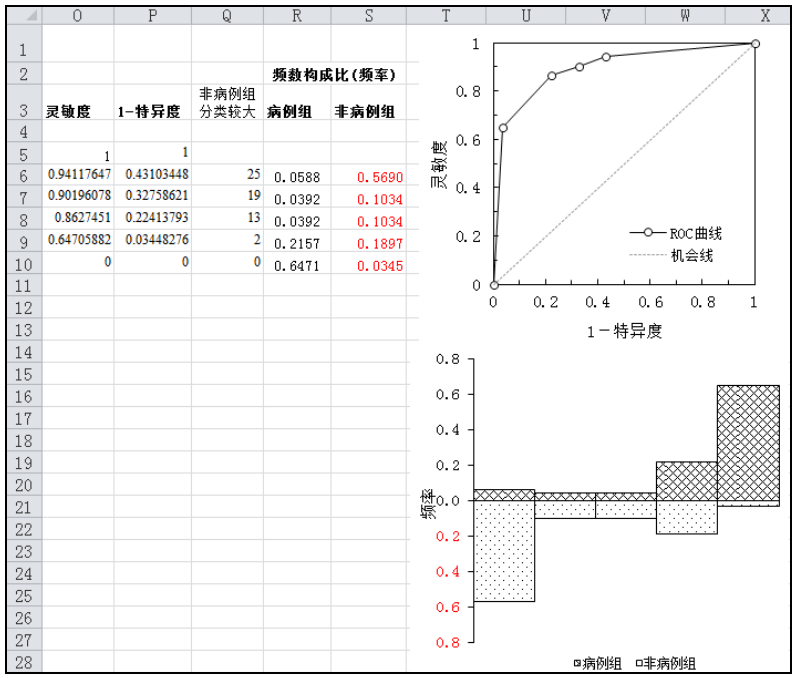
	A	B	C	D	E	F	G	H	I	J	K	L
1	ROC分析（值越大越有可能为阳性）											
2			金标准									
3	序号	分类或测量值	非病例组 X_n	病例组 X_a	病例组分类较大 Y_a	非病例组分类较小 Y_n	$X_aY_a+0.5X_aX_a$	$X_a(Y_a^2+Y_aX_a+X_a^2/3)$	$X_n(Y_a^2+Y_aX_a+X_a^2/3)$		最终结果	
4											分类数	5
5	合计		58	51			2642	123432.6667	142612.6667		n_a	51
6	1	1	33	3	48	0	1633.5	80883	1089		n_n	58
7	2	2	6	2	46	33	282	13256	2598		Q_1	0.818204316
8	3	3	6	2	44	39	270	12152	3534		Q_2	0.831250534
9	4	4	11	11	33	45	423.5	16415.66667	28163.66667			
10	5	5	2	33	0	56	33	726	107228		曲线下面积Az	0.893171062
11											面积标准误SE(Az)	0.031990408
12											95%CI下限值	0.830471014
13											96%CI上限值	0.955871109
14											Z	12.29027961
15											P	0

图 16-4 ROC 曲线下面积及其置信区间

结果分析：该例共有 5 个分类，正常组 n_n=58 例，异常组 n_a=51 例，ROC 曲线下面积 A_z=0.8932，接近 0.9，说明该放射医生采用 CT 诊断的能力较高，相应标准误 SE_{A_z} =0.0320，ROC 面积的总体 95%置信区间为（0.8305, 0.9559）。经 z 检验，得到 P 值远小于 0.05，说明曲线下面积 0.8932 与 0.5 之间差异有统计学意义（如图 16-4 所示）。ROC 曲线图及源数据频率直方图如图 16-5 所示。

第 N 列临界点是根据第 B 列“分类或测量值”获得，第一个值为第 B 列最小值减 1，最后一个值为第 B 列最大值加 1，其余为第 B 列相邻两值的平均值。以此临界值为截断点，大于该临界值的分类或测量值被划归为阳性，由此获得病例组真阳性率（即灵敏度）见第 O 列，非病例组假阳性率（即 1-特异度）见第 P 列，以这两列数据可绘制 ROC 曲线（如图 16-5 右上所示），该图的图表类型为“*XY* 散点图”，其中的机会线是在“选择数据→图

例项（系列）(S)”中添加（0,0）、（1,1）两个坐标点获得。第 R 列、S 列的病例组、非病例组频率分别由第 D 列、C 列的病例组、非病例组频数获得。图 16-5 右下方的源数据频率直方图（绘制方法参见第 7 章），实际上是病例组直方图与非病例组直方图绘制在同一个图中，该图形类似图 16-2，但为了避免两个直方图重叠难以区分，故将病例组直方图绘制在 0 点横坐标的上方，非病例组直方图绘制在 0 点横坐标的下方。细心读者会发现，非病例频率（第 S 列的频率）均为小于等于 0 的数，非 0 数字显示为红色，这是因为该列频率均加了负号，为了不显示负号，已修改数字格式为“数值”，“负数”显示为红色“1234.3210”。



	频数构成比(频率)				
	灵敏度	1-特异度	非病例组分类较大	病例组	非病例组
	1	1			
	0.94117647	0.43103448	25	0.0588	0.5690
	0.90196078	0.32758621	19	0.0392	0.1034
	0.8627451	0.22413793	13	0.0392	0.1034
	0.64705882	0.03448276	2	0.2157	0.1897
	0	0	0	0.6471	0.0345

图 16-5　ROC 曲线图及源数据频率直方图

温馨提示：该程序既适用于有序数据，也适用于连续数据，程序设定最大“分类或测量值”数量为 200。只要在单元格区域 B6:D205 范围内依次输入（也可以复制、粘贴）分类或测量值、非病例组频数、病例组频数，便可获得所有上述结果。必须注意，如果获得 ROC 曲线下面积小于 0.5，应检查病例组与非病例组数据是否恰好错误输入相反列；如果输入列没有错误，则可能是“值越小越有可能为阳性”，此时应选择光盘文件“16-3 小值阳性 ROC.xlsx”。

此外，ROC 曲线图及源数据频率直方图的源数据引用需要手工修改。修改前以密码“0”解除保护，方法是：选择“审阅→更改→撤销工作表保护”命令，输入密码即可。

16.2.2 连续资料 ROC 分析

例 16-3：Hans 等于 1985 年测量了 60 例有严重头外伤者 24 小时内脑脊髓液中肌氨酸激酶（creatine kinase-BB, CK-BB）值，其中 19 例最后完全康复，41 例预后较差或没有恢复（见光盘文件“16-4 ROC 曲线 0.xlsx”中的“data”工作表）。试根据此数据对患者预后（死亡、植物人、或严重残废）情况进行推断。

解：打开光盘文件“16-4 ROC 曲线 0.xlsx”中的“data”工作表，这一数据有 3 列，第 A 列为个体的编号“num”，第 B 列为每一个体相应的肌氨酸激酶“CK-BB”测量值，第 C 列为预后情况“outcome”，预后好（完全康复）记为 1，预后差（死亡等）记为 0。为了使用“16-2 ROC 曲线.xlsx”中的程序，需对此数据进行透视分析获得必要的临界截断值及对应的频数。方法如下：

① 将鼠标放在“data”工作表数据范围内，单击“插入→表→数据透视表/数据透视图→确定”，将“num”拖至数据区，“CK-BB”拖至行字段区，“outcome”拖至列字段区，双击“求和项:num”（单元格 A3），在弹出的“数据透视表字段”汇总方式中选择“数值计数”，单击“确定”。

② 选取单元格区域 A5:C57，将此数据复制、粘贴到新建工作簿中。

③ 在数据区被选定的情况下，单击“编辑→替换”命令，“替换为”后面空白处填“0”（查找内容不变，即空白），单击“全部替换”（这样将数据区 50 个空白格子替换为“0”），单击“确定”、“关闭”。

④ 将单元格区域 A1:C53 数据复制并选择性粘贴“数值”到光盘文件“16-4 ROC 曲线 0.xlsx”的“roc”工作表的 B6 单元格，便可获得如图 16-6 所示的结果。

	A	B	C	D	E	F	G	H	I	J	K	L
1	ROC分析（值越大越有可能为阳性）											
2			金标准									
3	序号	分类或测量值	非病例组 X_n	病例组 X_d	病例组分类较大 Y_d	非病例组分类较小 Y_n	$X_nY_n+0.5X_nX_d$	$X_n(Y_n^2+Y_nX_n+X_n^2/3)$	$X_d(Y_n^2+Y_dX_n+X_n^2/3)$		最终结果	
4											分类数	53
5	合计		19	41			645.5	22698	11006		n_d	41
6	1	6	1	0	41	0	41	1681	0		n_n	19
7	2	16	0	1	40	1	0	0	1		Q_1	0.710667209
8	3	17	1	0	40	1	40	1600	0		Q_2	0.743598406
9	4	23	1	0	40	2	40	1600	0			
10	5	27	1	0	40	3	40	1600	0		曲线下面积Az	0.828626444
11	6	40	1	0	40	4	40	1600	0		面积标准误SE(Az)	0.052282938
12	7	46	1	0	40	5	40	1600	0		95%CI下限值	0.726153769
13	8	60	1	1	39	6	39.5	1560.333333	42.33333333		95%CI上限值	0.93109912
14	9	70	1	0	39	7	39	1521	0		Z	6.285538972
15	10	76	0	2	37	8	0	0	128		P	3.26718E-10
16	11	80	0	1	36	8	0	0	64			
17	12	90	0	1	35	8	0	0	64			
18	13	96	1	0	35	8	35	1225	0			
19	14	100	2	0	35	9	70	2450	0			
20	15	120	0	1	34	11	0	0	121			
21	16	126	1	1	33	11	33.5	1122.333333	132.3333333			
22	17	136	1	0	33	12	33	1089	0			
23	18	140	0	1	32	13	0	0	169			
24	19	146	1	0	32	13	32	1024	0			

图 16-6 连续资料的 ROC 分析

结果分析：该例共有 53 个临界截断值，预后差 n_n=41 例，预后好 n_a=19 例，ROC 曲线下面积 A_z=0.8286，相应标准误 SE_{A_z} =0.0523，ROC 面积的总体 95%置信区间为（0.7262, 0.9311）。经 z 检验，得到 P 值远小于 0.05，说明曲线下面积 0.8286 与 0.5 之间差异有统计学意义。

	临界点	灵敏度	1-特异度	非病例组分类较大	频数构成比（频率） 病例组	非病例组
5	5	1	1			
6	11	1	0.94736842	18	0.0000	0.0526
7	16.5	0.97560976	0.94736842	18	0.0244	0.0000
8	20	0.97560976	0.89473684	17	0.0000	0.0526
9	25	0.97560976	0.84210526	16	0.0000	0.0526
10	33.5	0.97560976	0.78947368	15	0.0000	0.0526
11	43	0.97560976	0.73684211	14	0.0000	0.0526
12	53	0.97560976	0.68421053	13	0.0000	0.0526
13	65	0.95121951	0.63157895	12	0.0244	0.0526
14	73	0.95121951	0.57894737	11	0.0000	0.0526
15	78	0.90243902	0.57894737	11	0.0488	0.0000
16	85	0.87804878	0.57894737	11	0.0244	0.0000
17	93	0.85365854	0.57894737	11	0.0244	0.0000
18	98	0.85365854	0.52631579	10	0.0000	0.0526
19	110	0.85365854	0.42105263	8	0.0000	0.1053
20	123	0.82926829	0.42105263	8	0.0244	0.0000
21	131	0.80487805	0.36842105	7	0.0244	0.0526
22	138	0.80487805	0.31578947	6	0.0000	0.0526
23	143	0.7804878	0.31578947	6	0.0244	0.0000
24	149.5	0.7804878	0.26315789	5	0.0000	0.0526
25	154.5	0.75609756	0.26315789	5	0.0244	0.0000
26	169.5	0.73170732	0.26315789	5	0.0244	0.0000
27	188	0.70731707	0.26315789	5	0.0244	0.0000
28	196.5	0.68292683	0.26315789	5	0.0244	0.0000

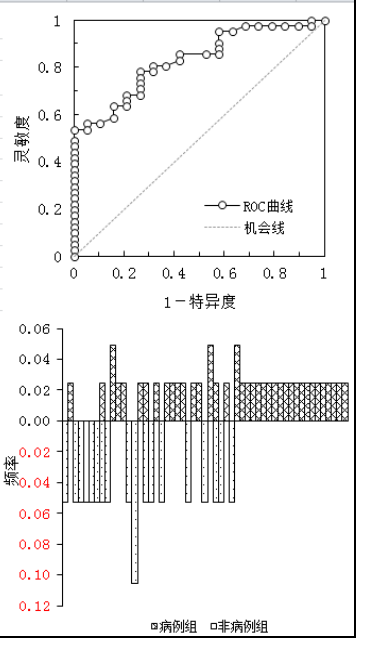


图 16-7　连续资料 ROC 曲线图及频率直方图

选择“审阅→更改→撤销工作表保护”命令，输入密码“0”，撤销工作表“roc”保护，在 ROC 曲线图上单击鼠标右键，修改源数据“ROC 曲线”系列的“X 值”、“Y 值”后的“10”为“58”，单击“确定”按钮，获得如图 16-7 右上角所示的 ROC 曲线图；在频率直方图上单击鼠标右键，修改数据源“病例组”系列以及“非病例组”系列中“值”后的“10”为“58”，单击“确定”按钮，再修改频率直方图的纵坐标格式中数字的“数值”小数位数为 2，这样便获得图 16-7 右下角的频率直方图。

16.2.3　值小为阳性的 ROC 分析

一般情况下，诊断试验分类值或测量值越大，定义为阳性的可能性越大，下面的实例恰好相反，即值越小越有可能为阳性。

例 16-4：采用骨髓诊断作为金标准，对 100 例患者进行诊断，其中 34 例确诊为缺

铁性贫血（异常组），其余 66 例确诊为非缺铁性贫血（正常组）。事先测得每个患者的红细胞平均容积（MCV）见光盘文件“16-3 小值阳性 ROC.xlsx”中的“data”工作表（资料来自 JR Beck, EK Shultz, Arch Pathol Lab Med，1986），试采用 ROC 分析评价 MCV 诊断缺铁性贫血的能力。

解：打开光盘文件“16-3 小值阳性 ROC.xlsx”中的“data”工作表，这一数据有 3 列，第 A、B、C 列分别为“编号”、“mcv 结果”、“骨髓诊断”，按例 16-3 中数据透视表方法（共 4 步）获得的数据，复制并选择性粘贴“数值”到光盘文件“16-3 小值阳性 ROC.xlsx”的“roc”工作表的 B6 单元格，便可获得如图 16-8 所示的结果。

	A	B	C	D	E	F	G	H	I	J	K	L
1	ROC分析（值越小越有可能为阳性）											
2			金标准									
3	序号	分类或测量值	非病例组	病例组	病例组分类较大	非病例组分类较小					最终结果	
4			X_n	X_a	Y_a	Y_n	$X_a Y_n+0.5X_a X_n$	$X_n(Y_a^2+Y_a X_n+X_n^2/3)$	$X_a(Y_a^2+Y_a X_n+X_a^2/3)$		分类数	38
5	合计		66	34			635	11417.33333	20313.33333		n_a	34
6	1	52	0	1	33	0	0	0	0		n_n	66
7	2	58	0	1	32	0	0	0	0		Q_1	0.149645241
8	3	60	1	0	32	0	32	1024	0		Q_2	0.137155872
9	4	62	0	1	31	1	0	0	1			
10	5	65	0	1	30	1	0	0	1		曲线下面积Az	0.717023173
11	6	66	1	0	30	1	30	900	0		面积标准误SE(Az)	0.052601184
12	7	67	0	1	29	2	0	0	4		95%CI下限值	0.613926747
13	8	68	1	1	28	2	28.5	812.3333333	6.333333333		96%CI上限值	0.820119599
14	9	69	1	1	27	3	27.5	756.3333333	12.33333333		Z	4.125822975
15	10	71	2	1	26	4	53	1404.666667	25.33333333		P	3.69411E-05
16	11	72	0	2	24	6	0	0	72			
17	12	73	1	2	22	6	23	529.3333333	84.66666667			
18	13	74	3	1	21	7	64.5	1387	73			
19	14	75	0	1	20	10	0	0	100			
20	15	76	1	1	19	10	19.5	380.3333333	110.3333333			
21	16	77	4	2	17	11	72	1297.333333	340.6666667			
22	17	78	2	1	16	15	33	544.6666667	256.3333333			
23	18	79	2	1	15	17	31	480.6666667	324.3333333			
24	19	80	2	2	13	19	28	392.6666667	800.6666667			

图 16-8 小值阳性的 ROC 分析

结果分析：该例共有 38 个临界截断值，正常值 n_n=66 例，异常组 n_a=34 例，ROC 曲线下面积 A_z=0.7170，相应标准误 SE_{A_z} =0.0526，ROC 面积的总体 95%置信区间为（0.6139, 0.8201）。经 z 检验，得到 P 值远小于 0.05，说明曲线下面积 0.7170 与 0.5 之间差异有统计学意义。

选择“审阅→更改→撤销工作表保护”命令，输入密码“0”，撤销工作表“ROC”保护，在 ROC 曲线图上单击鼠标右键，修改数据源“ROC 曲线”系列的“X 值”、“Y 值”后的最后两个数字为“43”，单击“确定”按钮，获得如图 16-9 右上角所示的 ROC 曲线图；在频率直方图上单击鼠标右键，修改源数据“病例组”系列以及“非病例组”系列中“值”后的最后两个数字为“43”，单击“确定”按钮，再修改频率直方图的纵坐标格式中数字的“数值”小数位数为 2，这样便获得如图 16-9 右下角所示的频率直方图。

	N	O	P	Q	R	S
1						
2					频数构成比（频率）	
3	临界点	灵敏度	1-特异度	非病例组分类较大	病例组	非病例组
4						
5	51	0	0			
6	55	0.02941176	0	66	0.0294	0.0000
7	59	0.05882353	0	66	0.0294	0.0000
8	61	0.05882353	0.01515152	65	0.0000	0.0152
9	63.5	0.08823529	0.01515152	65	0.0294	0.0000
10	65.5	0.11764706	0.01515152	65	0.0294	0.0000
11	66.5	0.11764706	0.03030303	64	0.0000	0.0152
12	67.5	0.14705882	0.03030303	64	0.0294	0.0000
13	68.5	0.17647059	0.04545455	63	0.0294	0.0152
14	70	0.20588235	0.06060606	62	0.0294	0.0152
15	71.5	0.23529412	0.09090909	60	0.0294	0.0303
16	72.5	0.29411765	0.09090909	60	0.0588	0.0000
17	73.5	0.35294118	0.10606061	59	0.0588	0.0152
18	74.5	0.38235294	0.15151515	56	0.0294	0.0455
19	75.5	0.41176471	0.15151515	56	0.0294	0.0000
20	76.5	0.44117647	0.16666667	55	0.0294	0.0152
21	77.5	0.5	0.22727273	51	0.0588	0.0606
22	78.5	0.52941176	0.25757576	49	0.0294	0.0303
23	79.5	0.55882353	0.28787879	47	0.0294	0.0303
24	80.5	0.61764706	0.31818182	45	0.0588	0.0303
25	81.5	0.70588235	0.36363636	42	0.0882	0.0455
26	82.5	0.73529412	0.39393939	40	0.0294	0.0303
27	83.5	0.76470588	0.5	33	0.0294	0.1061
28	84.5	0.79411765	0.56060606	29	0.0294	0.0606

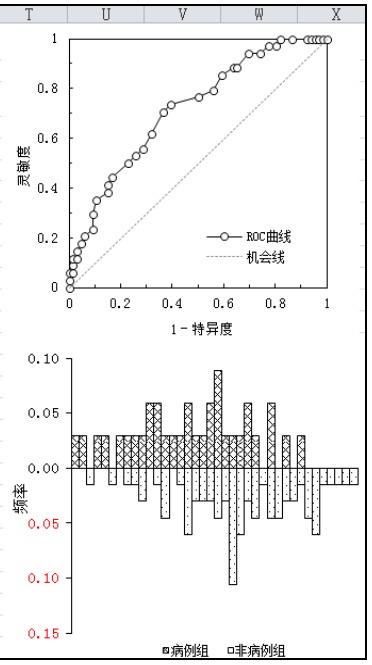


图 16-9　连续资料 ROC 曲线图及频率直方图

16.3　诊断试验评价的 Meta 分析

如果为了综合灵敏度或特异度等单一指标的信息，可按第 15 章的离散资料 Meta 分析方法（公式见表 15-4b）进行分析；如果每一研究均报道有灵敏度与特异度两个指标，可采用 16.3.1 节的“SROC 分析”；为了合并每一研究的 ROC 曲线下面积，可采用 16.3.2 节的“ROC 面积的 Meta 分析”。

16.3.1　SROC 分析

SROC 分析是综合接受者工作特征（Summary Receiver Operating Characteristic, SROC）分析的简称。SROC 线性回归模型为：

$$\hat{D} = a + b \times S \tag{16-5}$$

式中，D 为对数优势比，即

$$\begin{aligned} D &= \ln\frac{LR_+}{LR_-} = \ln\frac{Sen/(1-Sen)}{(1-Spe)/Spe} = \ln\frac{TP \times TN}{FP \times FN} \\ &= \ln OR = \text{logit}(Sen) - \text{logit}(1-Spe) \end{aligned} \tag{16-6}$$

其优势比在前面常用指标中已有计算，表示相对对照组，病例组中阳性结果的优势。*D* 值的大小反映诊断试验的“判别”能力。

S 与诊断临界截断点的选择有关，即

$$S = \ln LR_+ LR_- = \text{logit}(Sen) + \text{logit}(1 - Spe)$$
$$= \ln \frac{Sen/(1-Sen)}{Spe/(1-Spe)} = \ln \frac{TP \times FP}{TN \times FN} \quad (16\text{-}7)$$

当灵敏度（*Sen*）和特异度（*Spe*）相等时，*S*=0；当增加灵敏度或降低特异度（ROC 曲线从左下向右上伸展）时，*S*＞0；当降低灵敏度或增加特异度（ROC 曲线从右上向左下伸展）时，*S*＜0。*a* 为线性回归模型的截距项，是 *S*=0（即灵敏度等于特异度）时的对数优势比。*b* 为回归系数，反映对数优势比（*D*）依赖于阈值（*S*）的程度，如果回归系数与 0 的差异无统计学意义，则可采用截距项 *a* 作为多个原研究的公共对数优势比。

SROC 曲线的回归方程为：

$$Sen = \left[1 + e^{-a/(1-b)} \left(\frac{Spe}{1-Spe}\right)^{(1+b)/(1-b)}\right]^{-1} \quad (16\text{-}8)$$

采用 Sen^* 表示诊断试验准确度，Sen^* 为 *S*=0（即与机会对角线十字交叉的主对角线）与公式（16-8）联合求得的灵敏度 *Sen* 值，即

$$Sen^* = (1 + e^{-a/2})^{-1} \quad (16\text{-}9)$$

其对应的标准误为：

$$SE_{Sen^*} = \frac{S_a}{8[\cosh(a/4)]^2} \quad (16\text{-}10)$$

式中，S_a 为直线回归方程的截距项 *a* 对应的标准误，cosh 为双曲余弦函数。Sen^* 的 95% 置信区间为 $Sen^* \pm 1.96SE_{Sen^*}$。

每一个研究 *i* 相应的灵敏度 Sen_i 的置信区间为：

$$\left(\left[1 + e^{-a_L}\left(\frac{Spe_i}{1-Spe_i}\right)\right]^{-1}, \left[1 + e^{-a_U}\left(\frac{Spe_i}{1-Spe_i}\right)\right]^{-1}\right) \quad (16\text{-}11)$$

其中 a_L 与 a_U 为截距 *a* 的置信区间下、上限，95%置信区间下限为 $a_L = a - t_{0.05,\nu}S_a$，上限为 $a_U = a + t_{0.05,\nu}S_a$。

回归模型参数可采用一般线性回归模型（见第 10 章）、加权线性回归模型、稳健法等。加权线性回归模型的权重为：

$$W=\left(TP^{-1}+FP^{-1}+FN^{-1}+TN^{-1}\right)^{-1}$$
$$=\left(\frac{1}{(TP+FN)Sen(1-Sen)}+\frac{1}{(FP+TN)Spe(1-Spe)}\right)^{-1} \quad (16\text{-}12)$$

通过权重可以利用研究的样本含量信息，其他方法类似一般线性回归模型。1993 年，Moses、Shapiro 和 Littenberg 提出了 SROC 分析的稳健法，该方法先按诊断临界值 S 的大小，将散点（S,D）分成左、中、右三等分，记左边中位数散点为（S_1,D_1），右边中位数散点为（S_2, D_2），由公式

$$a=\frac{D_1S_2-D_2S_1}{S_2-S_1},\quad b=\frac{D_2-D_1}{S_2-S_1} \quad (16\text{-}13)$$

可获得公式（16-5）的 a、b 值，由公式（16-9）可获得合并灵敏度，由 exp(a)获得优势比。

例 16-5：Pap 试验是一种收集、制备和检查子宫颈脱落细胞，筛查宫颈癌的试验。Fahey、Irwig 和 Macaskill 按 Meta 分析要求收集了 59 篇 Pap 试验的真阳性数（*TP*）、假阳性数（*FP*）、假阴性数（*FN*）和真阴性数（*TN*）的值（如表 16-6 所示）（当有数据格子为 0 时，采用 0.5 替代），问 Pap 试验的诊断宫颈癌的性能如何？

表 16-6　59 篇Pap试验的四格表数据

编号 *i*	真阳性 *TP*	假阳性 *FP*	假阴性 *FN*	真阴性 *TN*	编号 *i*	真阳性 *TP*	假阳性 *FP*	假阴性 *FN*	真阴性 *TN*
1	8	3	23	84	31	11	1	1	2
2	31	3	43	14	32	23	50	10	44
3	70	12	121	25	33	65	13	42	13
4	65	10	6	6	34	1269	928	264	1084
5	20	3	19	4	35	223	22	74	83
6	35	92	20	156	36	154	30	20	237
7	39	8	111	270	37	6	2	12	81
8	567	117	140	157	38	7	4	3	4
9	25	37	11	18	39	12	5	11	60
10	38	28	17	37	40	348	41	212	103
11	45	35	15	48	41	8	4	11	34
12	71	87	10	306	42	12.5	2.5	6.5	0.5
13	4.5	0.5	36.5	5.5	43	95	9	2	1
14	2	2	3	21	44	40	18	20	19
15	5	9	3	182	45	71	13	20	18
16	38	21	7	62	46	1204	186	455	241

续表

编号 i	真阳性 TP	假阳性 FP	假阴性 FN	真阴性 TN	编号 i	真阳性 TP	假阳性 FP	假阴性 FN	真阴性 TN
17	4	2	16	31	47	6	20	51	27
18	87	13	12	9	48	35	9	12	12
19	15	3	65	15	49	10	31	5	32
20	41	1	61	29	50	3	5	3	15
21	76	12	11	12	51	118	40	44	183
22	10	4	48	174	52	13	3	82	17
23	28	11	28	77	53	38	14	13	62
24	3.5	0.5	5.5	1.5	54	14	25	67	291
25	79	26	13	182	55	12	14	6	12
26	61	20	27	35	56	238	52	2	16
27	62	20	16	49	57	111	44	20	39
28	284	31	68	68	58	491	165	250	701
29	66	25	20	44	59	48	16	38	31
30	40	43	12	47					

解：光盘文件“16-5 SROC 分析.xlsx”有“sroc”、“wls”两个工作表，其中“wls”工作表是一个辅助工作表，主要目的是为了便于采用加权最小二乘法获得加权线性回归模型的参数及其标准差，如果用户只为计算，可以不看该工作表。“sroc”工作表的数据输入（单元格区域 B5:E204）及部分之间的计算结果如图 16-10 所示。

	A	B	C	D	E	F	G	H	I	J	K
1	SROC分析										
2	编号	真阳性	假阳性	假阴性	真阴性	灵敏度	1-特异度	权重	$D=\frac{TP\times TN}{FP\times FN}$	$S=\ln\frac{TP\times FP}{TN\times FN}$	
3	i	TP	FP	FN	TN	$TP/(TP+FN)$	$FP/(FP+TN)$	$W=(TP^{-1}+FP^{-1}+FN^{-1}+TN^{-1})^{-1}$			
4	合计										
5	1	8	3	23	84	0.258064516	0.034482759	1.946599496	2.276151836	-4.388257184	
6	2	31	3	43	14	0.418918919	0.176470588	2.172610501	1.21323213	-1.867657952	
7	3	70	12	121	25	0.366492147	0.324324324	6.854784321	0.186673872	-1.281264479	
8	4	65	10	6	6	0.915492958	0.625	2.228571429	1.871802177	2.893453424	
9	5	20	3	19	4	0.512820513	0.428571429	1.457800512	0.338975367	-0.236388778	
10	6	35	92	20	156	0.636363636	0.370967742	10.4328321	1.087683218	0.031548358	
11	7	39	8	111	270	0.26	0.028776978	6.121684206	2.473011862	-4.564948973	
12	8	567	117	140	157	0.801980198	0.427007299	41.97631039	1.692788752	1.104645011	
13	9	25	37	11	18	0.694444444	0.672727273	4.684023427	0.100434397	1.541526707	
14	10	38	28	17	37	0.690909091	0.430769231	6.762210771	1.083086218	0.525659413	
15	11	45	35	15	48	0.75	0.421686747	7.230989957	1.414465238	0.782759339	
16	12	71	87	10	306	0.87654321	0.221374046	7.761159098	3.217771767	0.702417801	
17	13	4.5	0.5	36.5	5.5	0.109756098	0.083333333	0.411279308	0.304660409	-4.491130137	
18	14	2	2	3	21	0.4	0.086956522	0.724137931	1.945910149	-2.756840365	
19	15	5	9	3	182	0.625	0.047120419	1.538606049	3.517607734	-2.495956486	
20	16	38	21	7	62	0.844444444	0.253012048	4.293300937	2.774287958	0.609064063	
21	17	4	2	16	31	0.2	0.060606061	1.183770883	1.354545663	-4.127134385	
22	18	87	13	12	9	0.878787879	0.590909091	3.53529565	1.613276689	2.348726249	
23	19	15	2	65	15	0.1875	0.166666667	2.074468085	0.143100844	-3.075774981	

图 16-10　SROC 数据输入区与部分中间计算结果

每次研究的灵敏度、1-特异度、权重、D、S 的值及其计算公式分别位于图 16-10 的第 F、G、H、I、J 列。

三种模型拟合获得的主要计算结果如图 16-11 所示。一般线性模型与加权线性模型均包含回归模型斜率 b、截距 a、合并灵敏度 Sen^* 及其标准误，由于解释的必要后两个指标还给出了相应的 95%置信区间。由于稳健法不能获得相应指标的标准误，所以没有假设检验及其 95%置信区间。

	L	M	N	O	P	Q	R	S	T
1									
2	主要结果								
3	一般线性模型			加权线性模型				SROC稳健法	
4		结果			结果				结果
5	例数n	59		例数n	59			例数n	59
6	剩余标准差$S_{Y.X}$	1.144402708		剩余标准差$S_{Y.X}$	2.42966121				
7	决定系数R^2	0.001781358		决定系数R^2	0.00081304			$b=\frac{D_2-D_1}{S_2-S_1}$	0.001122
8									
9	b	0.02008492		b	-0.0150811				
10	S_b	0.06297525		S_b	0.07002666			$a=\frac{D_1S_2-D_2S_1}{S_2-S_1}$	1.591351
11	t_b值	0.318933557		t_b值	-0.2153627				
12	P值	0.750942731		P值	0.83025327			优势比	4.910378
13									
14	a	1.589981759		a	1.71978668				
15	S_a	0.151103433		S_a	0.09982849			合并准确度Sen^*	0.689049
16	t_a值	10.5224728		t_a值	17.227413				
17	P值	5.62533E-15		P值	2.899E-24				
18	a_L(95%CI)	1.287402354		a_L(95%CI)	1.51988357				
19	a_U(95%CI)	1.892561163		a_U(95%CI)	1.91968978				
20	优势比	4.903659479		优势比	5.58333727				
21									
22	合并准确度Sen^*	0.688901963		合并准确度Sen^*	0.70263837				
23	$SE(Sen^*)$	0.016191945		$SE(Sen^*)$	0.01042897				
24	Sen*(95%CI下限)	0.657165751		Sen*(95%CI下限)	0.68219759				
25	Sen*(96%CI上限)	0.720638176		Sen*(96%CI上限)	0.72307915				
26	Z值	42.54596665		Z值	67.3737238				
27	P值	0		P值	0				

图 16-11　三种模型的计算结果

如果以原数据获得的“1-特异度”（见第 G 列）为横轴，根据公式（16-8）可计算纵轴“灵敏度”（见图 16-12 的第 X、AB、AI 列），绘制的 SROC 曲线见三个图的平行短线所形成的曲线（一般线性模型和加权线性模型位于置信区间中间）。一般线性模型和加权线性模型的 95%置信区间在图中显示为空心圆点围成的曲线，根据公式（16-11）获得。稳健法因为无法获得标准误，所以无 95%置信区间，其图中散在的空心散点为每一研究原数据“灵敏度”（见第 F 列）与“1-特异度”所对应的坐标点。

为了检验是否具有发表偏倚，以稳健法获得的各研究灵敏度（见第 AI 列）为横坐标，各研究样本含量（$N_i=TP_i+FP_i+FN_i+TN_i$）为纵坐标绘制的漏斗图，如图 16-13 所示。

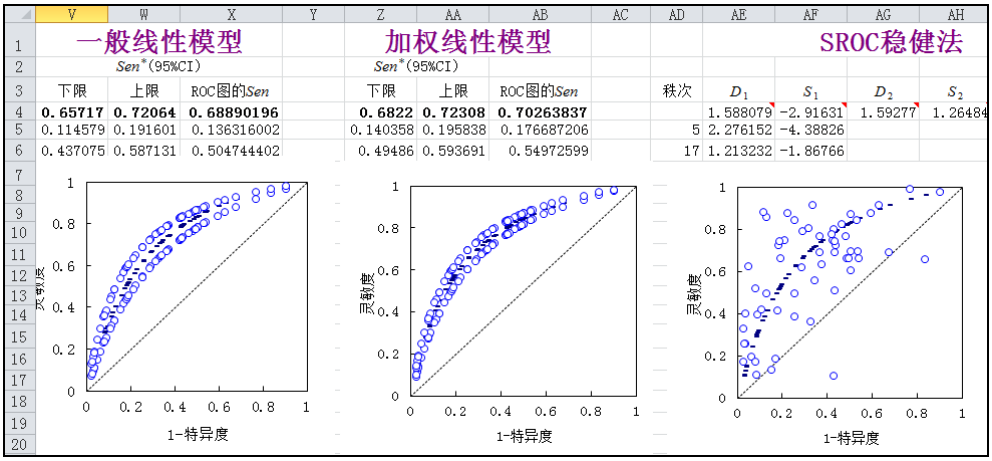


图 16-12 三种模型的图形表达

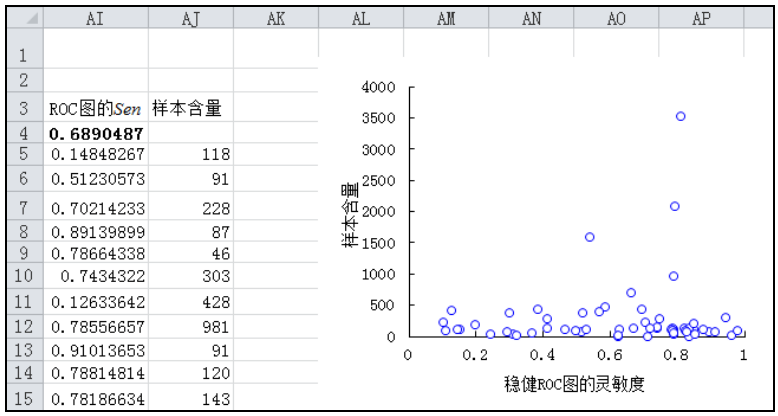


图 16-13 稳健法漏斗图

结果分析：由图 16-11 可见，三种方法获得的结果基本一致，即 59 个研究的合并灵敏度 Sen^*约为 0.7。这个值的解释类似曲线面积，即值在 0.5～0.7 之间表示诊断准确度较低，0.7～0.9 之间表示诊断准确度为中等，0.9 以上表示诊断准确度较高。本例合并灵敏度 Sen^* 约为 0.7，表明 Pap 试验诊断宫颈癌的性能中等偏低。由图 16-12 可见，加权线性模型的 95% 置信区间的范围较狭窄，表明其估计精确度更高。图 16-13 所显示的漏斗图基本上是对称的倒置漏斗形状，所以可以认为出版偏倚可以忽略。三个模型的其他主要结果如表 16-7 所示，b、S_b、a、S_a 分别为线性回归模型的斜率、斜率标准误、截距、截距标准误，Sen^*、SE_{Sen^*} 分别为合并 SROC 灵敏度、合并灵敏度标准误。

表 16-7　三个回归模型计算的诊断试验准确度

方　法	b	S_b	a	S_a	a 的 95%CI	优势比	Sen^*	SE_{Sen^*}	Sen^*的 95%CI
一般线性模型	0.020	0.063	1.590	0.151	1.287～1.893	4.904	0.689	0.016	0.657～0.721
加权线性模型	-0.015	0.070	1.720	0.100	1.520～1.920	5.585	0.703	0.010	0.682～0.723
稳健法	0.001	—	1.591	—	—	4.911	0.689	—	—

温馨提示：该程序最大可输入 200 个研究的数据。只需在“sroc”工作表的单元格区域 B5:E204 中相应位置依次输入已知的各研究真阳性数 *TP*、假阳性数 *FP*、假阴性数 *FN*、真阴性数 *TN* 便可获得所有结果。但必须注意，如果某研究的 *TP*、*FP*、*FN*、*TN* 中有“0”值应将此替换为“0.5”，以避免计算中分母为 0 的情况。

程序中图形的源数据无法自动改变，需要根据源数据所占单元格区域情况进行修改。图 16-12 右侧图形显示了实际各研究“1-特异度”与“灵敏度”的散点及其稳健法 SROC 曲线，同样也可以将一般线性模型 SROC 曲线和加权 SROC 曲线绘制在该图中。

16.3.2　ROC 面积的 Meta 分析

如果诊断试验结果为 ROC 曲线下面积或部分面积，那么可采用类似第 15 章所介绍的方法进行分析，这种方法假定 ROC 面积服从正态分布，合并效应指标即 ROC 曲线下面积（记为 A_z），相应标准误记为 SE_{A_z}，95%置信区间为 $A_z \pm 1.96SE_{A_z}$，其他计算公式如表 16-8 所示。

表 16-8　ROC曲线下面积的Meta分析计算公式

方　法	权　重	合并效应值	合并标准误
固定效应	$w_i = 1/SE_{A_{zi}}^2$	$A_z = \dfrac{\Sigma w_i A_{zi}}{\Sigma w_i}$	$SE_{A_z} = \sqrt{1/\Sigma w_i}$
齐性检验	$Q = \Sigma w_i \left(A_{zi} - A_z\right)^2$，　$\nu = k - 1$		
随机效应	$w_i^* = 1/(1/w_i + SE_\tau^2)$	$A_{zR} = \dfrac{\Sigma w_i^* A_{zi}}{\Sigma w_i^*}$	$SE_{A_{zR}} = \sqrt{1/\Sigma w_i^*}$
	其中 $SE_\tau^2 = \max\left(0, \dfrac{Q-k+1}{\Sigma w_i - \Sigma w_i^2/\Sigma w_i}\right)$		

例 16-6：地塞米松抑制试验（Dexamethasone Suppression Test, DST）是一种评价垂体肾上腺调节失衡的试验，可用于诊断与区别各种精神疾患，如精神忧郁症、精神分裂症和躁狂症等。Mossman 和 Somaza 于 1989 年收集的 7 个 DST 研究如表 16-9 所示，请利用该数据对 DST 试验作出评价。

表 16-9 7 个地塞米松抑制试验

编号 i	A_{zi}	$SE(A_{zi})$	患者人数	
			病 例 组	对 照 组
1	0.789	0.057	34	33
2	0.724	0.025	215	152
3	0.851	0.028	119	79
4	0.876	0.029	54	41
5	0.782	0.102	52	49
6	0.702	0.056	65	31
7	0.652	0.038	111	77

解：光盘文件“16-6 ROC 面积 Meta 分析.xlsx”有“Az”、“置信区间图”、“漏斗图”3 个工作表，其中“Az”工作表分为数据输入区（第 B、C、D、E 列）、Meta 分析计算结果区（第 G、H、I、J、K 列）（见图 16-14）及中间结果计算区（如图 16-15 所示）。

	A	B	C	D	E	F	G	H	I	J	K
1	ROC曲线下面积的Meta分析										
2	请输入数据！								分析结果		
3	研究编号	ROC面积	标准误	患者人数				研究个数k	7		
4	i	A_{zi}	$SE(A_{zi})$	病例组 n_{di}	对照组 n_{ni}			齐性检验Q值	35.7136648	自由度	6
5	1	0.789	0.057	34	33			P值	3.13331E-06	拒绝H0，齐性不成立	
6	2	0.724	0.025	215	152						
7	3	0.851	0.028	119	79				ROC面积	标准误	Z值
8	4	0.876	0.029	54	41			固定效应	0.780831818	0.01350874	20.7889024
9	5	0.782	0.102	52	49			随机效应	0.769711859	0.03599609	7.4928092
10	6	0.702	0.056	65	31						
11	7	0.652	0.038	111	77				下限	上限	
12								固定效应总体ROC面积的95%CI	0.754354693	0.80730894	
13								随机效应总体ROC面积的95%CI	0.699159516	0.8402642	
14											
15									最终结果		
16									ROC面积		
17								随机效应	0.769711859		
18									下限	上限	
19								随机效应总体ROC面积的95%CI	0.699159516	0.8402642	
20								Z值	7.4928092	P值	6.75E-14
21										合并效应值有统计学意义	

图 16-14 Meta 分析数据及计算结果

光盘文件“16-6 ROC 面积 Meta 分析.xlsx”的“置信区间图”工作表界面如图 16-16 所示。其图形中间竖线为无诊断价值的曲线下面积即 A_z=0.5，其他与前面类似。光盘文件“16-6 ROC 面积 Meta 分析.xlsx”的“漏斗图”工作表以各研究的 ROC 曲线下面积为横轴，样本含量为纵轴绘制的一种散点图，如图 16-17 所示。可能例数较少，倒置漏斗图左下角略有空缺。

	M	N	O	P	Q	R	S	T	U	V	W
1											$SE_z^2=\max(0,\frac{Q-k+1}{\Sigma w_i-\Sigma w_i^2/\Sigma w_i})$
2	研究	固定效应				齐性检验			随机效应		
3	编号	权重		合并效应指标		Q			权重		合并效应指标
4	i	$w_i=1/SE_{Ai}^2$	w_iA_{ai}	$A_z=\Sigma w_iA_{ai}/\Sigma w_i$		$Q=\Sigma w_i(A_{ai}-A_z)^2$		w_i^2	$w_i^*=1/(1/w_i+S\hat{E}_z^2)$	$w_i^*A_{ai}$	$A_{zR}=\Sigma w_i^*A_{ai}/\Sigma w_i^*$
5	合计	5479.873062	4278.859243	0.780831818		35.7136648		6286030.707	771.7724355	594.042396	0.769711859
6	1	307.7870114	242.843952			0.020535304		94732.84438	98.94224818	78.0654338	
7	2	1600	1158.4	标准误		5.167768782		2560000	133.6379212	96.753855	标准误
8	3	1275.510204	1085.459184	$SE_{A_z}=\sqrt{1/\Sigma w_i}$		6.280068654		1626926.281	130.8574068	111.359653	$SE_{A_{zR}}=\sqrt{1/\Sigma w_i^*}$
9	4	1189.060642	1041.617122	0.013508737		10.76930196		1413865.211	129.8885845	113.7824	0.035996093
10	5	96.11687812	75.16339869			0.000131166		9238.45426	57.93103039	45.3020658	
11	6	318.877551	223.8520408	Z值		1.981650339		101682.8925	100.0609773	70.2428061	Z值
12	7	692.5207756	451.5235457	20.78890238		11.4942086		479585.0247	120.454267	78.5361821	7.492809204

图 16-15　中间计算结果

	A	B	C	D	E	F	G
1	总体ROC曲线下面积的95%置信区间						
2	研究编号	ROC面积	95%CI下限	96%CI上限	区间下半	区间上半	权重百分率(%)
3	0	0.7697119	0.699159516	0.8402642	0.07055234	0.0705523	
4	1	0.789	0.67728	0.90072	0.11172	0.11172	5.62
5	2	0.724	0.675	0.773	0.049	0.049	29.20
6	3	0.851	0.79612	0.90588	0.05488	0.05488	23.28
7	4	0.876	0.81916	0.93284	0.05684	0.05684	21.70
8	5	0.782	0.58208	0.98192	0.19992	0.19992	1.75
9	6	0.702	0.59224	0.81176	0.10976	0.10976	5.82
10	7	0.652	0.57752	0.72648	0.07448	0.07448	12.64

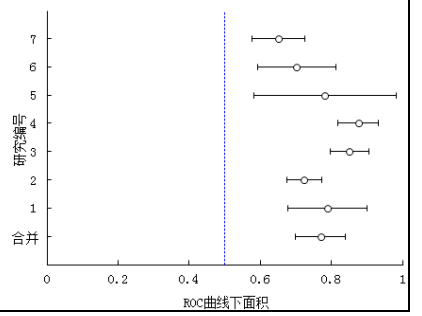


图 16-16　ROC 曲线下面积的置信区间图

	A	B	C	D	E	F	G	H	I	J
1	漏斗图（funnel plot）									
2	研究编号	ROC面积	样本含量							
3	1	0.789	67							
4	2	0.724	367							
5	3	0.851	198							
6	4	0.876	95							
7	5	0.782	101							
8	6	0.702	96							
9	7	0.652	188							

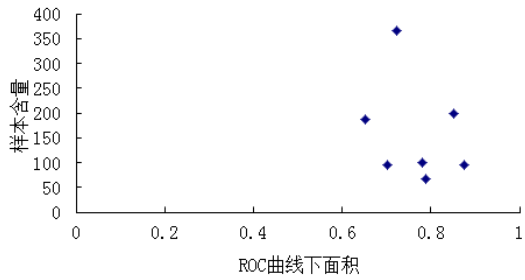


图 16-17　ROC 面积的漏斗图

结果分析：由图 16-14 可见，齐性检验 Q=35.7137，在自由度为 6 的卡方分布中，得到 P 值远小于 0.05，表明研究间 ROC 曲线下面积的差异较大，应采用随机效应模型的结果。随机效应模型获得的合并效应值为 A_z =0.7697，95%置信区间为（0.6991,0.8402），与

A_z=0.5 相比较，差异具有统计学意义（z=7.4928，P=0.0000）。

温馨提示：该程序最大可输入 100 个研究的数据。只需在“Az”工作表的单元格区域 B5:E104 中相应位置输入已知数据便可获得所有结果。但必须注意，如果研究数不同（即不为 7 个研究），则 95%置信区间图以及漏斗图的相应源数据做必要改变。如数据输入区原已有数据存在，最好先清除旧数据再输入新数据。

16.4 样本含量估计

光盘文件“16-7 诊断试验样本含量.xlsx”共有 6 个工作表，分别对灵敏度、特异度、ROC 曲线下面积的置信区间或假设检验所需的样本含量进行估计。

16.4.1 灵敏度与特异度的置信区间估计

灵敏度（sensitivity，*Sen*）或特异度（specificity，*Spe*）的双侧$100(1-\alpha)\%$置信区间的估计公式分别为：

$$Sen \pm z_{\alpha/2} SE_{Sen} \text{ 或 } Spe \pm z_{\alpha/2} SE_{Spe}$$

其中α为检验水准，$z_{\alpha/2}$为标准正态分布双侧分位数，如α=0.05，即$100(1-\alpha)\%$=95%置信区间的$z_{\alpha/2} = z_{0.025} = 1.96$，$SE_{Sen}$与$SE_{Spe}$分别表示灵敏度与特异度的标准误。

$$SE_{Sen} = \sqrt{\frac{Sen(1-Sen)}{n_{有病}}}$$

$$SE_{Spe} = \sqrt{\frac{Spe(1-Spe)}{n_{无病}}}$$

其中$n_{有病}$、$n_{无病}$分别为“有病”组、“无病”组研究对象的人数。估计以上置信区间所需要的样本含量分别为：

$$n_{有病} = \frac{z_{\alpha/2}^2 Sen(1-Sen)}{L^2} \quad (16\text{-}14)$$

$$n_{无病} = \frac{z_{\alpha/2}^2 Spe(1-Spe)}{L^2} \quad (16\text{-}15)$$

其中 L 为期望置信区间宽度的一半，即为$z_{\alpha/2}SE_{Sen}$或$z_{\alpha/2}SE_{Spe}$。

例 16-7：某研究灵敏度估计值为 0.80，特异度估计值为 0.75，期望的置信区间宽度的一半 L = 0.05，要估计真实灵敏度或特异度 95%置信区间，所需要的样本含量至少为多少？

解：打开光盘文件“16-7 诊断试验样本含量.xlsx”的“1.灵敏度与特异度区间估计”工作表，将已知数据 α =0.05、*Sen* =0.8、*Spe*=0.75 和 δ =0.05 输入到单元格区域 C4:C7，可分别获得真实灵敏度双侧 95%置信区间对应的病例组例数 $n_{有病}$ =246，真实特异度双侧 95%置信区间对应的对照组例数 $n_{无病}$ =289，两组合计样本含量为 $n_{合并}$ =535。如果只需获得单侧 95%置信限，则对应样本含量分别为 $n_{有病}$ =174、$n_{无病}$ =203、$n_{合并}$ =377。注意：灵敏度只对应病例组样本含量估计，特异度只对应对照组样本含量估计（如图 16-18 所示）。

灵敏度与特异度置信区间估计的样本含量

请输入有关数据				所需最低的样本含量结果		
					双侧	单侧
1.检验水准	α	0.05		$n_{合计}$	535	377
2.灵敏度	*Sen*	0.8		$n_{有病}$	246	174
3.特异度	*Spe*	0.75		$n_{无病}$	289	203
4.容许误差	δ	0.05				

计算公式与中间结果

	双侧	单侧	
$n_{有病}=\frac{Z_{\alpha/2}^2 Sen(1-Sen)}{\delta^2}$	245.8533645	173.154781	$n_{有病}=\frac{Z_{\alpha}^2 Sen(1-Sen)}{\delta^2}$
$n_{无病}=\frac{Z_{\alpha/2}^2 Spe(1-Spe)}{\delta^2}$	288.1094116	202.915759	$n_{无病}=\frac{Z_{\alpha}^2 Spe(1-Spe)}{\delta^2}$

图 16-18　灵敏度与特异度的置信区间估计

16.4.2　ROC 曲线下面积的置信区间估计

ROC 曲线下面积的双侧 $100(1-\alpha)\%$ 置信区间估计公式为 $A_z \pm z_{\alpha/2}SE_{A_z}$，其中 A_z 为 ROC 曲线下面积，SE_{Az} 为 A_z 的标准误。样本含量为：

$$n_{有病}=(1+c)\frac{z_{\alpha/2}^2 V(A_z)}{L^2},\quad c=\frac{n_{无病}}{n_{有病}} \tag{16-16}$$

其中 c 表示研究对象中“无病”组与“有病”组患者的比值，总样本含量为 $n_{合并}=(1+c)n_{有病}$。

当采用 Hanley 和 McNeil（1982）非参数方法计算 ROC 曲线下面积时，

$$V(Az)=\frac{Az}{k(2-Az)}+2Az^2/(1+Az)-Az^2(1/k+1) \tag{16-17}$$

如果假定“无病”组与“有病”组对应总体为双正态分布，则采用 Obuchowski（1994）提出的公式：

$$V(Az)=(0.0099\times e^{-a^2/2})\times[(5a^2+8)+(a^2+8)/k] \tag{16-18}$$

其中 a 为定义双正态分布 ROC 曲线

$$\Phi^{-1}(Sen)=a+b\Phi^{-1}(1-Spe),\quad 0\leqslant(1-Spe)\leqslant 1$$

的参数，即

$$a=\Phi^{-1}(Az)\times 1.414 \tag{16-19}$$

其中 $\Phi^{-1}(Az)=\Phi^{-1}(0.975)=1.96$。通常情况下令双正态分布 ROC 曲线的另一个参数 b= 1.0。

例 16-8：如果 ROC 曲线下面积估计值为 0.85，期望的置信区间宽度的一半 $L=0.1$，“无病”组与“有病”组患者的比值 c =0.176，问估计真实 ROC 曲线下面积的 95%置信区间需要多少样本含量？

解：打开光盘文件“16-7 诊断试验样本含量.xlsx”的“2.ROC 面积区间估计”工作表，将已知数据 α =0.05、A_z =0.85、c =0.176 和 δ =0.1 输入到单元格区域 C4:C7 中，可获得参数法双侧 95%置信区间对应的病例组 $n_{有病}$ =100，对照组 $n_{无病}$ =18，两组合计样本含量为 $n_{合并}$ =118；非参数法双侧 95%置信区间对应的病例组 $n_{有病}$ =59，对照组 $n_{无病}$ =11，两组合计样本含量为 $n_{合并}$ =70。如果只需获得单侧 95%置信限，则对应参数法样本含量分别为 $n_{有病}$ =70、$n_{无病}$ =13、$n_{合并}$ =83；非参数法样本含量分别为 $n_{有病}$ =42、$n_{无病}$ =8、$n_{合并}$ =50（如图 16-19 所示）。

ROC曲线下面积置信区间估计的样本含量

			所需最低的样本含量结果				
请输入有关数据			双正态假定参数法		非参数法		
			双侧	单侧	双侧	单侧	
1. 检验水准	α	0.05					
2. ROC曲线下面积	A_z	0.85	$n_{合计}$	118	83	70	50
3. 样本量比值	$c=n_{无病}/n_{有病}$	0.176	$n_{有病}$	100	70	59	42
4. 容许误差	δ	0.1	$n_{无病}$	18	13	11	8

计算公式与中间结果

		双正态假定参数法		非参数法	
		双侧	单侧	双侧	单侧
$n_{合计}=n_{无病}+n_{有病}$	合计	116.7169334	82.2038579	69.15121151	48.70327039
双侧：$n_{有病}=\frac{Z^2_{\alpha/2}V(Az)}{\delta^2}$	$n_{有病}$	99.24909299	69.90123971	58.8020506	41.41434557
$n_{无病}=cn_{有病}$	$n_{无病}$	17.46784037	12.30261819	10.34916091	7.28892482

双正态假定下的参数法：$V(Az)=(0.0099\times e^{-2Z^2_{1-Az}/2})\times\left[(10Z^2_{1-Az}+8)+(2Z^2_{1-Az}+8)/c\right]$

非参数法：$V(Az)=\frac{Az}{c(2-Az)}+2Az^2/(1+Az)-Az^2(1/c+1)$

图 16-19　ROC 曲线下面积的置信区间估计

16.4.3　固定 FPR 的灵敏度置信区间估计

在给定 ROC 曲线两参数 a、b 条件下，固定假阳性率 FPR=1−Spe = e 处的灵敏度为：

$$Sen_{\mathrm{FPR}=e}=\Phi\left(a+b\Phi^{-1}(e)\right) \tag{16-20}$$

因为 $Sen_{\mathrm{FPR}=e}$ 不服从正态，需要进行正态变换，即

$$z\left(Sen_{\mathrm{FPR}=e}\right)=a+b\Phi^{-1}(e) \tag{16-21}$$

所给定的置信区间一半 δ 对应的是灵敏度 $Sen_{\mathrm{FPR}=e}$，为了获得 $z\left(Sen_{\mathrm{FPR}=e}\right)$ 对应的置信区间一半 δ'，需计算 $Sen'=Sen_{\mathrm{FPR}=e}\pm\delta$，$a'=\Phi^{-1}(Sen')-b\Phi^{-1}(e)$，$z\left(Sen'\right)=a'+b\Phi^{-1}(e)$，$\delta'=z\left(Sen'\right)_{\text{上限}}-z\left(Sen'\right)_{\text{下限}}$。病例组样本含量为：

$$n_{\text{有病}}=\frac{z_{\alpha/2}^2\left[(1+b^2/c+a^2/2)+\left(\Phi^{-1}(e)\right)^2\left(b^2(1+c)/(2c)\right)\right]}{\delta'^2} \tag{16-22}$$

其中 c 表示研究对象中“无病”组与“有病”组患者的比值，总样本含量为 $n_{\text{合并}}=(1+c)n_{\text{有病}}$。

例 16-9：某研究拟在 FPR = 0.10 处，计算灵敏度在 ±0.10 范围的 95%置信区间所需的样本含量，假定对照组与病例组样本含量比值为 0.176，ROC 曲线两参数 a=1.03，b=0.5。请估计所需的最低样本含量。

解：打开光盘文件“16-7 诊断试验样本含量.xlsx”的“3.固定 FPR 的灵敏度区间估计”工作表，将已知数据 α=0.05，e=0.1，c=0.176、a=1.03，b=0.5 和 δ=0.1 输入到单元格区域 C4:C9，可获得双侧 95%置信区间对应的病例组 $n_{\text{有病}}$=220，对照组 $n_{\text{无病}}$=39，两组合计样本含量为 $n_{\text{合并}}$=259；如果只需获得单侧 95%置信限，则对应样本含量分别为 $n_{\text{有病}}$=155、$n_{\text{无病}}$=28、$n_{\text{合并}}$=183（如图 16-20 所示）。

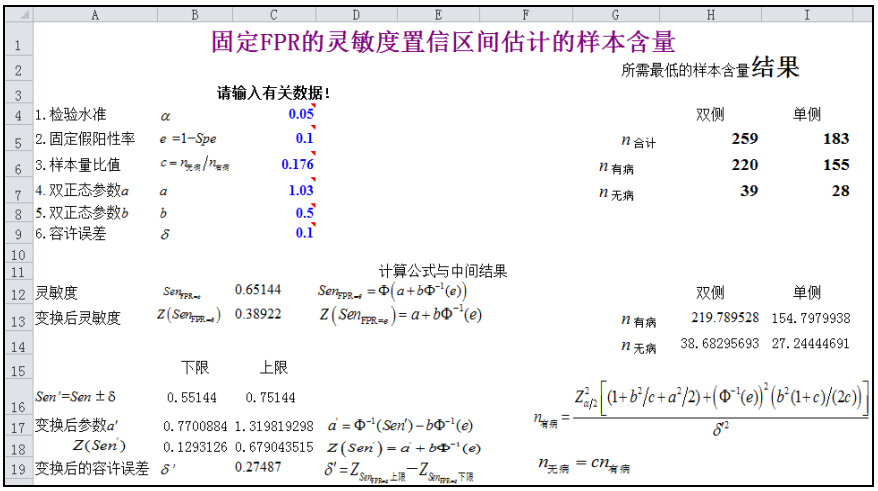


图 16-20　固定 FPR 处的灵敏度置信区间估计

16.4.4　单个准确度指标的假设检验

检验灵敏度、特异度、ROC 曲线下面积等单个准确度指标是否等于某一特定值，可以

采用本节的计算方法获得样本含量估计值。

在通常情况下，想检验 ROC 曲线面积是否不等于 0.5（灵敏度、特异度不为 0），零假设和备择假设分别为：

$$H_0：\theta=\theta_0=0.50，\ H_1：\theta\neq0.50=\theta_1$$

计算样本含量时，必须说明备择假设下的θ_1值，此时病例组患者样本含量的计算公式为：

$$n_{有病}=\frac{\left(z_{\alpha/2}V(\theta_0)+z_{\beta}V(\theta_1)\right)^2}{(\theta_0-\theta_1)^2} \tag{16-23}$$

其中β为II型错误的概率，z_{β}与$z_{\alpha/2}$的意义相同，都是标准正态分布的分位数。不同的是α可以取单侧和双侧（但一般取双侧），而β只取单侧，通常单侧$z_{0.05}=1.645$，$z_{0.10}=1.282$。$V(\theta_0)$为零假设下准确度指标θ的方差函数，$V(\theta_1)$为备择假设下θ的方差函数，即

$$V(\theta)=(0.0099\times e^{-2z_{\theta}^2/2})\times\left[(10z_{\theta}^2+8)+(2z_{\theta}^2+8)/c\right] \tag{16-24}$$

例 16-10：假定 H_1：Az_1=0.8，期望的置信区间宽度的一半为 0.1，“无病”与“有病”患者的比值c=1，以 I 型错误的概率为 0.05，检验功效$(1-\beta)$=90%，要检验真实 ROC 曲线下面积Az_1=0.8 是否有意义（即$Az_1>0.5$），问所需要的样本含量至少为多少？

解：打开光盘文件“16-7 诊断试验样本含量.xlsx”的“4.单个准确度指标比较”工作表，将已知数据α=0.05、$1-\beta$=0.9、θ_0=0.5、θ_1=0.8、c=1 和δ=0.1 输入到单元格区域 C4:C9，可获得检验Az_1=0.8 是否有意义所对应的病例组$n_{有病}$=17，对照组$n_{无病}$=17，两组合计样本含量为$n_{合并}$=34；如果只需获得单侧检验，则对应样本含量分别为$n_{有病}$=14、$n_{无病}$=14、$n_{合并}$=28（如图 16-21 所示）。

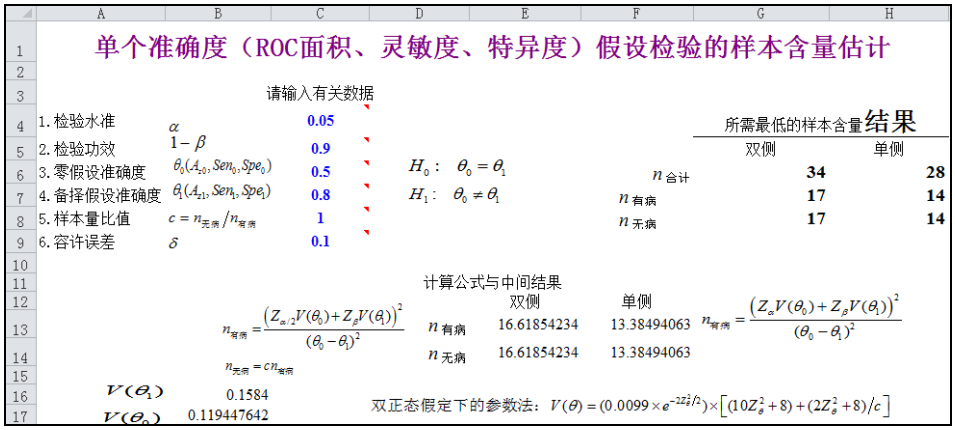


图 16-21 单个准确度指标的假设检验

16.4.5 两灵敏度或特异度的比较

灵敏度和特异度的样本含量估计方法类似，下面以灵敏度为例说明样本含量的估计方法。如果零假设和备择假设分别为：

$$H_0：Sen_1 = Sen_2，\ H_1：Sen_1 \neq Sen_2 \tag{16-25}$$

其中 Sen_1 与 Sen_2 为两个诊断试验的准确度，为了计算样本含量，需要在备择假设下指定 Sen_1 与 Sen_2。

样本含量的估计不仅依赖于是何种准确度指标，而且也依赖于研究是采用配对还是非配对设计。配对设计与非配对设计的四格表格式为：

诊断方法1 \ 诊断方法2	+	−	
+	a	b	$a+b$
−	c	d	
合计	$a+c$		n

	+	−	
诊断方法1	a_1	b_1	a_1+b_1
诊断方法2	a_2	b_2	a_2+b_2

配对设计　　非配对设计

配对设计的灵敏度样本含量计算公式（beam，1992）为：

$$n_{有病} = \frac{\left(z_{\alpha/2}\sqrt{\psi} + z_{\beta}\sqrt{(\psi - (Sen_2 - Sen_1)^2)}\right)^2}{(Sen_2 - Sen_1)^2} \tag{16-26}$$

其中 $z_{\alpha/2}$ 是标准正态分布的双侧分位数；α 为 I 型错误率；z_{β} 是标准正态分布的单侧分位数；β 为II型错误率；ψ 为两个诊断方法的不一致概率，其最小值为

$$(Sen_2 - Sen_1) \tag{16-27}$$

最大值为：

$$Sen_1(1 - Sen_2) + Sen_2(1 - Sen_1) \tag{16-28}$$

对于非配对两个诊断试验灵敏度的比较，样本含量估计公式（beam，1992）为：

$$n_{有病} = \frac{A\left(1 + \sqrt{1 + 4(Sen_2 - Sen_1)/A}\right)^2}{4(Sen_2 - Sen_1)^2} \tag{16-29}$$

其中：

$$A = \left(z_{\alpha/2}\sqrt{\frac{(Sen_1 + Sen_2)(2 - Sen_1 - Sen_2)}{2}} + z_{\beta}\sqrt{Sen_1(1 - Sen_1) + Sen_2(1 - Sen_2)}\right)^2$$

两个诊断试验特异度比较的公式与此类似，只不过所计算得到的样本含量为“无病”组的样本含量。

例 16-11：怀疑两个诊断试验之间灵敏度之差为 0.10，诊断方法 1 的灵敏度 Sen_1=0.7，诊断方法 2 的 Sen_1=0.8。在 I 型错误率为 5%，功效为 80%的前提条件下，试确定配对研究设计与非配对研究设计所需的样本含量至少各是多少？

解：打开光盘文件“16-7 诊断试验样本含量.xlsx”的“5.两灵敏度比较”工作表，将已知数据 α=0.05，$1-\beta$=0.8，Sen_1=0.7、Sen_2=0.8 输入到单元格区域 C4:C7，可获得配对设计对于配对设计所需的双侧样本含量（即对子数）范围为 77 至 296（分别对应最小不一致率 0.1 和最大不一致率 0.38）；单侧（将 $Z_{\alpha/2}$ 采用 Z_α 取代）所需的样本含量范围为 60 至 233。对于非配对设计，需要 313 名“有病”患者，两组共需 626 名“有病”患者。单侧所需“有病”患者为 251，两组共需要 502 例“有病”患者（如图 16-22 所示）。

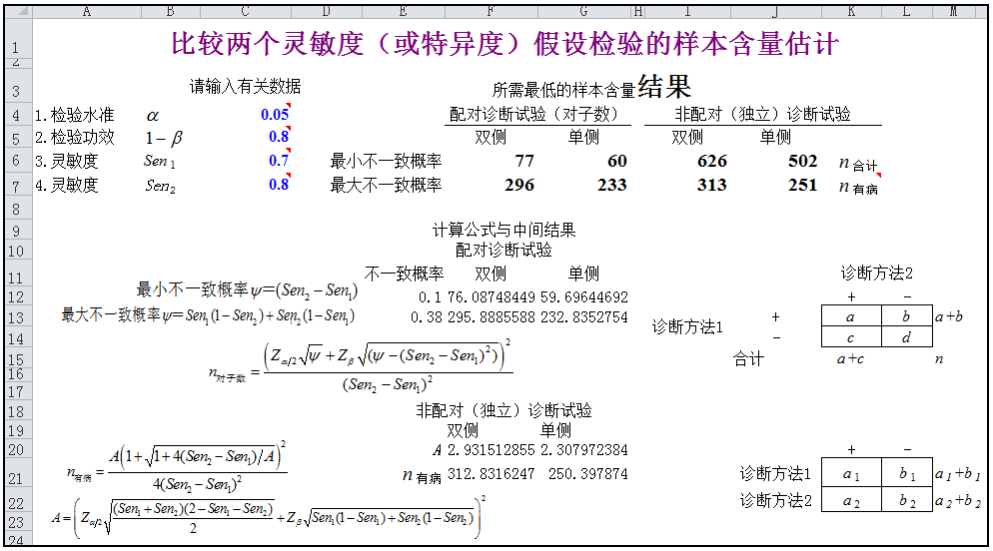


图 16-22 两个灵敏度（特异度）假设检验

16.4.6 ROC 曲线下面积的比较

两个 ROC 曲线下面积比较的“有病”组样本含量的计算公式为：

$$n_{有病}=\frac{\left[z_{\alpha/2}\sqrt{V_0(A_{z2}-A_{z1})}+z_{\beta}\sqrt{V_1(A_{z2}-A_{z1})}\right]^2}{(A_{z2}-A_{z1})^2} \tag{16-30}$$

其中 $V_0(A_{z2}-A_{z1})$ 为 H_0：$A_{z1}=A_{z2}$ 对应的方差函数，$V_1(A_{z2}-A_{z1})$ 为 H_1：$A_{z1}\neq A_{z2}$ 对应的方差函数，它们的计算公式为：

$$V_i(A_{z2}-A_{z1})=V(A_{z1})+V(A_{z2})-2C(A_{z1},A_{z2}) \tag{16-31}$$

根据双正态分布原理及泰勒展开式，其中$V(A_{z1})$、$V(A_{z2})$为：

$$V(A_z)=(0.0099\times e^{-a^2/2})\times[(5a^2+8)+(a^2+8)/c] \tag{16-32}$$

$$C(A_{z1},A_{z2})=\frac{e^{[-(a_1^2+a_2^2)/4]}}{12.5664}\times\left(\rho_D+\frac{\rho_N}{k}+\frac{\rho_D^2a_1a_2}{2}\right)+\frac{e^{[-(a_1^2+a_2^2)/4]}}{50.2655}\times\left(\frac{a_1a_2(\rho_N^2+k\rho_D^2)}{2c}\right)-\frac{e^{[-(a_1^2+a_2^2)/4]}}{25.1327}\times\left(\rho_D^2a_1a_2\right) \tag{16-33}$$

ρ_D与ρ_N分别为两个比较诊断试验之间“有病”组的相关系数与“无病”组的相关系数，c为“无病”组与“有病”组之间的样本含量之比。a_i、b_i为双正态 ROC 曲线的两个参数，a_i与b_i的关系式为：

$$a_i=\sqrt{1+b_i^2}\Phi^{-1}(A_{zi}) \tag{16-34}$$

以上是配对设计所需的“有病”组所需要的样本含量，配对设计共需要患者$n_{有病}(1+c)$。如果两个诊断试验之间独立，为非配对设计，则$C(A_{z1},A_{z2})$为 0，每一诊断方法所需的样本含量为$n_{有病}(1+c)$，两个诊断试验共需要$2n_{有病}(1+c)$。

例 16-12：诊断方法 1 与诊断方法 2 的曲线下面积分别为A_{z1}=0.80 与A_{z2}=0.90，零假设下A_{z1}=A_{z2}=0.90。令c=2，定义两条 ROC 曲线的参数$b_1=b_2=1$，问在$\alpha=0.05$，检验效能$1-\beta=0.8$，两个比较诊断试验之间“有病”组的相关系数与“无病”组的相关系数ρ_D及ρ_N分别为 0.7 时，需要多少样本含量？

解：打开光盘文件“16-7 诊断试样样本含量.xlsx”的“6.两 ROC 面积比较”工作表，将已知数据α=0.05，$1-\beta$=0.8，c=2、H_1：A_{z1}=0.8、H_1：A_{z2}=0.9、H_0：$A_{z1}=A_{z2}$=0.9、ρ_D=0.7、ρ_N=0.7、b=1，输入到单元格区域 C4:C12，比较诊断试验 1 与诊断试验 2 的 ROC 曲线下面积需要 50 例“有病”患者，126 例“无病”患者，共需约 189 例观察个体。单侧所需的样本含量：“有病”组患者为 50 例，“无病”组患者为 100 例，共需约 150 例观察个体。

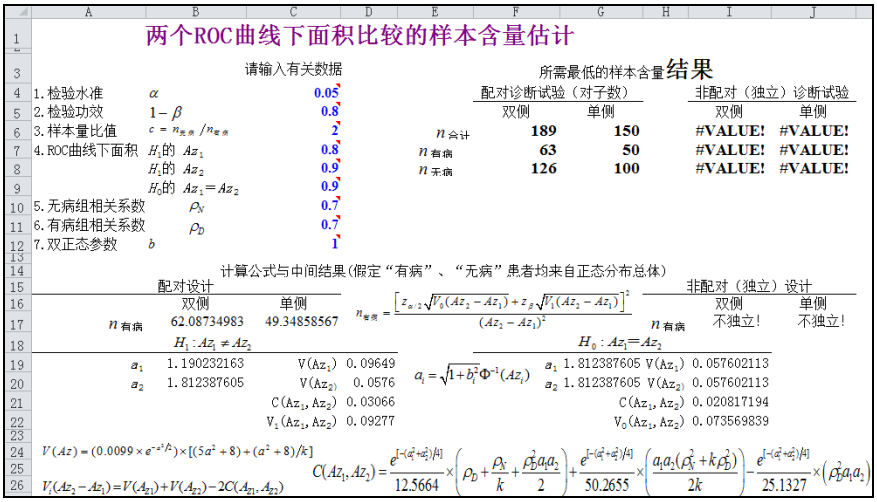
如果两个诊断试验之间独立，即$\rho_D=\rho_N$=0，$C(Az_1,Az_2)$=0，则每一诊断方法双侧需要“有病”患者 100 例，“无病”患者 200 例，每一诊断方法共计需要 300 例，两个诊断试验共计需要 600 例。单侧所需的样本含量：每一诊断方法“有病”患者为 79 例，“无病”患者需要 158 例，两个诊断试验共计需要 474 例（如图 16-23 所示）。

16.5 Kappa 值估计

在 16.1 节介绍了诊断试验评价四格表的 Kappa 统计量计算，实际上 Kappa 统计量可用于检查两个等级以上的两个观测一致性程度。Kappa 值的理论取值在 0～1 范围，Kappa 值为 0～0.4 时说明一致性程度不理想；Kappa 值≥0.75 时说明具有较好的一致性。Kappa 值κ

计算公式为：

$$\kappa=\frac{p_O-p_E}{1-p_E}=\frac{\dfrac{\text{对角实际观测一致数之和}}{\text{总例数}}-\dfrac{\text{对角理论期望一致数之和}}{\text{总例数}}}{1-\dfrac{\text{对角理论期望一致数之和}}{\text{总例数}}} \tag{16-35}$$



两个ROC曲线下面积比较的样本含量估计

			请输入有关数据	所需最低的样本含量结果	配对诊断试验（对子数）		非配对（独立）诊断试验	
					双侧	单侧	双侧	单侧
1. 检验水准	α		0.05	$n_{合计}$	189	150	#VALUE!	#VALUE!
2. 检验功效	$1-\beta$		0.8	$n_{有病}$	63	50	#VALUE!	#VALUE!
3. 样本量比值	$c=n_{无病}/n_{有病}$		2	$n_{无病}$	126	100	#VALUE!	#VALUE!
4. ROC曲线下面积	H_1的	Az_1	0.8					
	H_1的	Az_2	0.9					
	H_0的	$Az_1=Az_2$	0.9					
5. 无病组相关系数	ρ_Y		0.7					
6. 有病组相关系数	ρ_D		0.7					
7. 双正态参数	b		1					

计算公式与中间结果（假定“有病”、“无病”患者均来自正态分布总体）

配对设计

	双侧	单侧
$n_{有病}$	62.08734983	49.34858567

$$n_{有病}=\frac{\left[z_{\alpha/2}\sqrt{V_0(Az_2-Az_1)}+z_{\beta}\sqrt{V_1(Az_2-Az_1)}\right]^2}{(Az_2-Az_1)^2}$$

非配对（独立）设计

	双侧	单侧
$n_{有病}$	不独立!	不独立!

$H_1: Az_1 \neq Az_2$

a_1	1.190232163	$V(Az_1)$	0.09649
a_2	1.812387605	$V(Az_2)$	0.0576
		$C(Az_1, Az_2)$	0.03066
		$V_1(Az_1, Az_2)$	0.09277

$$a_i=\sqrt{1+b_i^2}\,\Phi^{-1}(Az_i)$$

$H_0: Az_1=Az_2$

a_1	1.812387605	$V(Az_1)$	0.057602113
a_2	1.812387605	$V(Az_2)$	0.057602113
		$C(Az_1, Az_2)$	0.020817194
		$V_0(Az_1, Az_2)$	0.073569839

$$V(Az)=(0.0099\times e^{-a^2/2})\times[(5a^2+8)+(a^2+8)/k]$$

$$V_i(Az_2-Az_1)=V(Az_1)+V(Az_2)-2C(Az_1,Az_2)$$

$$C(Az_1,Az_2)=\frac{e^{[-(a_1^2+a_2^2)/4]}}{12.5664}\times\left(\rho_D+\frac{\rho_Y}{k}+\frac{\rho_D^2a_1a_2}{2}\right)+\frac{e^{[-(a_1^2+a_2^2)/4]}}{50.2655}\times\left(\frac{a_1a_2(\rho_Y^2+k\rho_D^2)}{2k}\right)-\frac{e^{[-(a_1^2+a_2^2)/4]}}{25.1327}\times\left(\rho_D^2a_1a_2\right)$$

图 16-23 两个 ROC 曲线下面积的假设检验

其标准误为：

$$SE_\kappa=\frac{1}{(1-p_E)\sqrt{n}}\sqrt{p_E+p_E^2-\frac{\sum R_iC_i(R_i+C_i)}{n^3}} \tag{16-36}$$

式中，p_O 表示实际观察一致率，理论期望一致率 $p_E=\sum R_iC_i/N^2$，n 为总例数，R_i、C_i 分别为第 i 等级的行、列合计。Kappa 值 95%置信区间为 $\kappa\pm1.96SE_\kappa$，$z=\kappa/SE_\kappa$。

更为精确的标准误（SAS 软件计算公式）为：

$$SE_\kappa=\sqrt{n\left\{\frac{\left(\sum f_{ii}\right)(n-\sum f_{ii})}{\left(n^2-\sum R_iC_i\right)^2}+\frac{2(n-\sum f_{ii})\left(2\sum f_{ii}\sum R_iC_i-n\sum f_{ii}(R_i+C_i)\right)}{\left(n^2-\sum R_iC_i\right)^3}+\frac{(n-\sum f_{ii})^2\left[n\sum_{i,j}f_{ij}(R_j+C_i)^2-4\left(\sum R_iC_i\right)^2\right]}{\left(n^2-\sum R_iC_i\right)^4}\right\}} \tag{16-37}$$

式中，f_{ii} 为第 i 行与第 i 列对应格子频数，n 为总例数，R_i、C_i 分别为第 i 等级的行、列合计。Kappa 值 95%置信区间为 $\kappa \pm 1.96SE_\kappa$，$z=\kappa/SE_\kappa$。

例 16-13：两名放射科医生对 200 名棉屑沉着病可疑者的诊断分类结果如表 16-10 所示，试计算两个医生诊断结果一致性的 Kappa 值，并进行统计学检验。

表 16-10　200 名棉屑沉着病可疑者的诊断分类结果

第一人检查	第二人检查			合　计
	正　常	Ⅰ期	Ⅱ期	
正常	78	5	0	83
Ⅰ期	6	56	13	75
Ⅱ期	0	10	32	42
合计	84	71	45	200

解：打开光盘文件“16-8 Kappa 统计量修正.xlsx”，将诊断分类数据依次输入到单元格区域 B5:K14，便可自动获得如图 16-24 所示的结果。

结果分析：由图 16-24 可见，两个医生的诊断一致性 Kappa 值为 0.7366，标准误为 0.0405，95%置信区间为（0.6571,0.8160），说明两个医生诊断一致性较高。

温馨提示：该程序最大可处理 10 个分类等级的一致性。只要将等级数据依次输入到单元格区域 B5:K14，程序便自动获得所需结果。程序根据单元格区域 B5:K14 的数据情况自动编第 A 列和第 3 行序号。

采用上述修正的公式计算获得的 Kappa 值及其标准误与 SAS 软件（见光盘文件“16-8 kappa.sas”）计算的“Simple Kappa”一致。

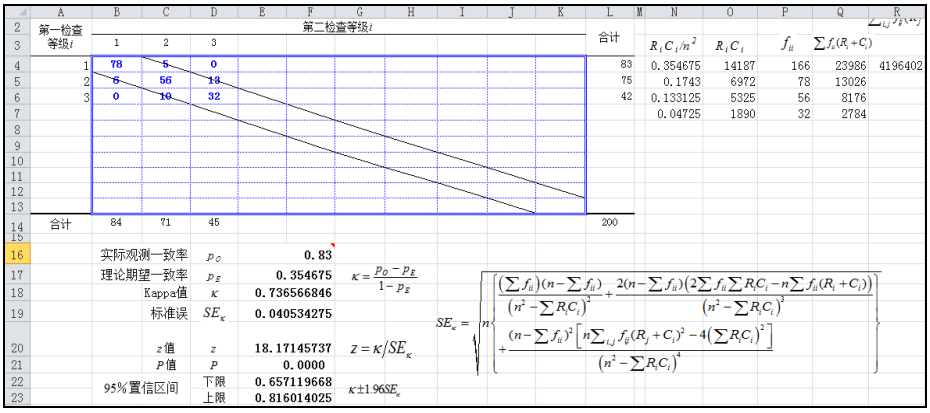


图 16-24　Kappa 统计量及其统计学检验

16.6 Bland-Altman 法

Bland-Altman 法可用于两定量变量的一致性检验，该方法最初由 Bland JM 和 Altman DG 于 1986 年提出。其基本思想是计算出两种测量结果的一致性界限（limits of agreement），并用图形直观地反映这个一致性界限。最后结合临床实际情况，得出两种测量方法是否具有一致性的结论。

记两方法测定结果差值的均数为 $\bar{d}$，差值的标准差为 S_d，如果差值服从正态分布，则 95 %的差值应该位于 $\bar{d} \pm 1.96S_d$ 范围。这一区间就是 95%一致性界限。如果两种测量结果的差值位于一致性界限内在临床上可以接受，则可认为这两种方法具有较好的一致性。

以两方法测量每个对象的均数为横轴，相应差值为纵轴，可绘制 Bland-Altman 一致性界限图。如图 15-25 中可有四条水平线，即 95%一致性上、下界限，均数线（中间较粗的线）和 $d = 0$ 的线（虚线）。

根据图中 95 %一致性界限外的数据点数、一致性界限内的最大差值、以及临床上的可接受程度，可对两种方法的一致性程度做出评价。

例 16-14：如图 16-25 左侧所示的调查员 1、调查员 2 的数据（也见光盘文件“16-9 Bland Altman 法.xlsx”），试采用 Bland-Altman 法进行一致性分析。

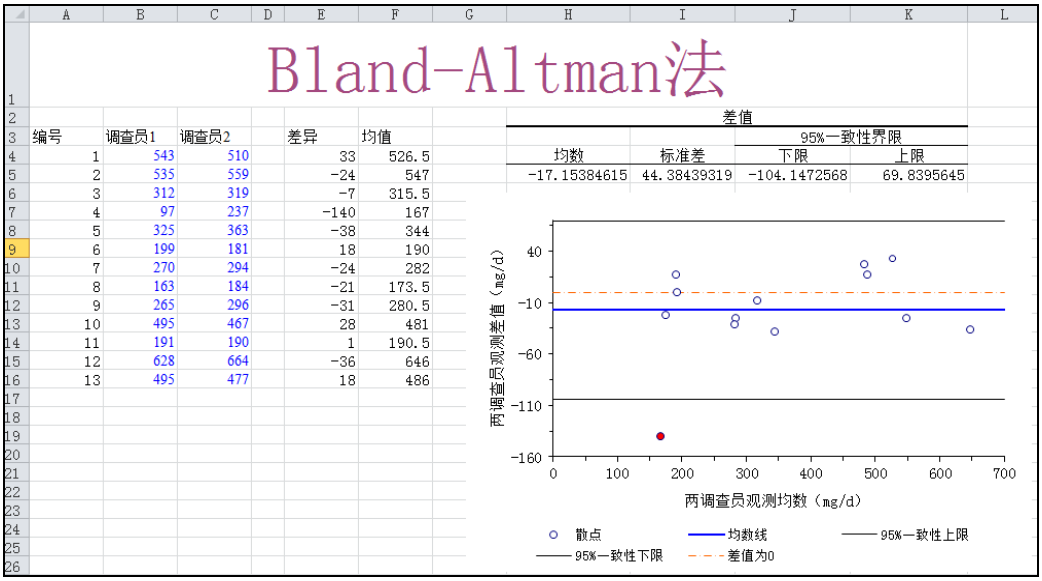


图 16-25 两调查员调查 13 名妇女食物中每日钙摄入量的 Bland Altman 图

解：先对 13 名妇女的每一位计算相应的两调查员的差值与均数（如图 16-25 所示的第 E、第 F 列），差值对应的均数、标准差分别为 $\bar{d} = -17.5$、$S_d = 44.38$，所以 95%一致性界限为（–104.15, 69.84）。

结合以上的计算结果，以均数为横轴，差值为纵轴，绘制的 Bland-Altman 图如图 16-25 所示。

由图 16-25 可见，只有 7.69%（1/ 13）的点在 95 % 一致性界限以外（实心点表示）；在一致性界限范围内，差值的最大值为 33，最小值为–38，差异不明显。也就是说调查员观测结果相差 33～–38，从实际考虑，这种差异可以接受，因此这两个观察员的观测结果具有较好的一致性。

温馨提示：该程序最大可处理 100 个“调查员”对子的一致性。只要将两个“调查员”数据输入到单元格区域 B4:C103，程序便自动获得所需结果。图形也会发生相应的改变，只需在图形上单击鼠标右键，单击快捷菜单上的“选择数据→编辑”命令，修改图形的横抽最大值、最小值、及均数与差值所在的单元格区域等，便可获得所需要的图形。

电脑实验篇

第 17 章　概率分布

第 18 章　随机抽样与抽样误差

第 19 章　假设检验的概率分布与统计用表

第 20 章　其他电脑实验与置信区间估计

第 17 章

概 率 分 布

概率（probability）是某事件发生的可能性（likelihood）或机率（chance），有关概率的基本概念见附录 B。概率分布（probability distribution）是进行统计学推论的基础和依据，一般分为连续概率分布（continuous probability distribution）与离散概率分布（discrete probability distribution）两大类。连续分布的数据理论上是无限可分的，而离散分布的数据只是整数，两相邻整数间没有小数。本章首先介绍几个重要的、常用的分布，如连续分布的正态分布、离散分布的二项分布和 Poisson 分布，然后介绍其他的概率分布。

17.1 正态分布

正态分布（normal distribution）也称为高斯（Gaussian）分布，是最基本的连续分布，常简记为 $N(\mu,\sigma^2)$。μ、σ 两个参数可决定一个正态分布曲线，μ 为总体均数，是位置参数，反映了最高峰（即均数）距原点的位置，当 μ 逐渐变小时，曲线往左移动；当 μ 逐渐变大时，曲线往右移动。σ 为总体标准差，是形状参数，反映了曲线的“胖”、“瘦”程度，σ 越小曲线越瘦高，表明数据越集中；σ 越大曲线越矮胖，表明数据越分散。其概率密度函数（probability density function, PDF）为：

$$f(X)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(X-\mu)^2}{2\sigma^2}},\quad -\infty<X<+\infty \qquad (17\text{-}1)$$

其中 $f(X)$ 为曲线的纵坐标；X 为曲线的横坐标，即实际取值的变量。π 为圆周率 3.141…，e 为自然对数的底 2.718…。

例 7-1（第 7 章第 1 个例子）的 130 名正常成年男子红细胞计数值（即 X 值）就是典型的正态分布例子。计算获得的均数为 4.7（10^{12}/L），标准差为 0.6（10^{12}/L），如果假定正常成年男子红细胞计数值的总体均数 μ =4.7，总体标准差 σ =0.6，则其分布可记为 $N(4.7,0.6^2)$。

17.1.1 采用公式制作正态分布电脑实验

例 17-1：光盘文件 17-1 有三个正态分布分别是 $N(0,\ 1^2)$、$N(-1,\ 0.9^2)$、$N(1,\ 1.1^2)$，试绘制与比较这三个正态分布的曲线。

解：将已知的总体均数 μ 和总体标准差 σ 代入公式（17-1），改变 X 值可获得不同的 $f(X)$。以 X 为横坐标，$f(X)$ 为纵坐标，可绘制出正态分布曲线。

为了获得绘制正态分布曲线的数据，可按下列具体步骤进行操作。

① 按图 17-1 所示键入单元格区域 A2:D2、G2:I2、F3:F4 中的文本。

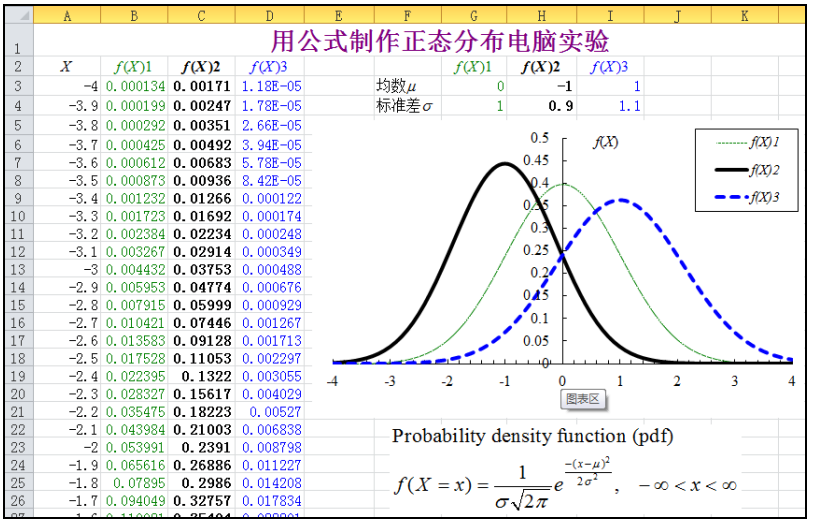
② 为了使绘出的曲线两端不至于太长或太短，X 的最小值与最大值分别为（均数±4 标准差），即–4～+4（均数为 0，标准差为 1）。X 的步长以（标准差/10）为宜，即 0.1。单元格 A3 的值为–4，单元格 A4 的值为–3.9，选取单元格 A3:A4，向下拖放填充至单元格 A83。

③ 单击单元格 B3，在此单元格（根据公式（17-1））中键入 Excel 公式“=1/(G$4*SQRT(2*PI()))*EXP(($A3–G$3)^2/(2*G$4^2)*(–1))”，按 Enter 键。将单元格 B3 中的公式向右拖放至 D3，然后向下拖放填充至单元格 D83。

对于产生好的数据，可采用图表向导中的 *XY* 散点图绘制曲线（可参见第 6 章“统计

图表”）。具体操作步骤如下：

① 单击“插入→图表→散点图（图标为 ）”，在下拉菜单中选择“所有图表类型”，在弹出的插入图表界面中，选取“*XY* 散点图”子图表类型中的“带平滑线的散点图”（如图 17-2 所示）。



用公式制作正态分布电脑实验

X	f(X)1	f(X)2	f(X)3
-4	0.000134	0.00171	1.18E-05
-3.9	0.000199	0.00247	1.78E-05
-3.8	0.000292	0.00351	2.66E-05
-3.7	0.000425	0.00492	3.94E-05
-3.6	0.000612	0.00683	5.78E-05
-3.5	0.000873	0.00936	8.42E-05
-3.4	0.001232	0.01266	0.000122
-3.3	0.001723	0.01692	0.000174
-3.2	0.002384	0.02234	0.000248
-3.1	0.003267	0.02914	0.000349
-3	0.004432	0.03753	0.000488
-2.9	0.005953	0.04774	0.000676
-2.8	0.007915	0.05999	0.000929
-2.7	0.010421	0.07446	0.001267
-2.6	0.013583	0.09128	0.001713
-2.5	0.017528	0.11053	0.002297
-2.4	0.022395	0.1322	0.003055
-2.3	0.028327	0.15617	0.004029
-2.2	0.035475	0.18223	0.00527
-2.1	0.043984	0.21003	0.006838
-2	0.053991	0.2391	0.008798
-1.9	0.065616	0.26886	0.011227
-1.8	0.07895	0.2986	0.014208
-1.7	0.094049	0.32757	0.017834

均数μ	f(X)1	f(X)2	f(X)3
标准差σ	0	-1	1
	1	0.9	1.1

图 17-1　三条正态分布曲线的部分数据

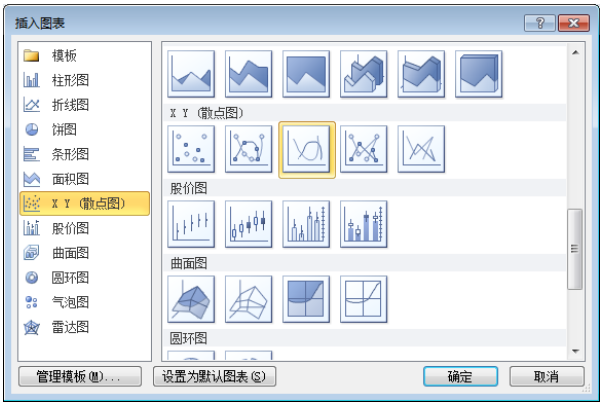


图 17-2　XY 散点图的图表向导

② 单击“确定”按钮，在生成的空白框处单击鼠标右键，在弹出的选项中单击“选择数据”，图表数据区域采用鼠标选择单元格区域“$A$2:$D$83”，自动生成“图例项（系列）”和“水平（分类）轴标签”，（如图 17-3 所示），单击“确定”按钮。

③ 为了使获得的图形美观好看，可以单击鼠标右键清除图域的背景色、网格线，改变

横轴、纵轴刻度，添加横轴、纵轴标题，改变曲线颜色和样式，调整图例位置等。



图 17-3　在数据区域选择数据

④为了便于辨认，将三条曲线的数据、文字、曲线改变成相同颜色。修改后的曲线如图 17-1 所示。

温馨提示　用户只需改变某曲线的均数与标准差，相应曲线将随之改变，这对于理解正态分布曲线的性质十分有用。通过改变某曲线的均数与标准差，可以发现正态分布的如下性质：

①曲线在 $X=\mu$ 处最高；
②曲线关于 $X=\mu$ 对称；
③随 μ 值不同，曲线位置不同，故称 μ 为位置参数；
④σ 大时，曲线“矮而胖”；σ 小时，曲线“瘦而高”，故称 σ 为形状参数。

公式（17-1）对应的 Excel 函数格式为：

NORMDIST（变量值 X，均数 μ，标准差 σ，0）

利用此 Excel 函数，可更简单地获得图 17-1 的第 B、C、D 列。

电脑实验

反复修改曲线 2（$f(X)2$）、曲线 3（$f(X)3$）的均数（单元格 H3、I3）或标准差（单元格 H4、I4），观察曲线 2、曲线 3 相对于标准正态曲线 $f(X)1$ 的变化情况，体会上述正态分布的性质。

17.1.2　正态曲线下的面积

正态分布曲线下的面积，即变量落在曲线下某区间的概率。实际工作中经常需要知道正态曲线下、横轴上方一定区间的正态曲线下面积占总面积的比例有多大，或者说需要知

道某个变量值落在某区间的概率是多少。

例 17-2：见光盘文件 17-2 某研究获得某地某年婴儿出生体重均数为 3250 克，标准差为 360 克，试采用正态分布原理估计该地当年体重低于 2400 克的婴儿占多大比例。

解：现已知 X=2400，μ=3250，σ=360，求正态曲线下横轴坐标从 $-\infty$～2400 的面积。

求解曲线下面积，需要对公式（17-1）概率密度函数取积分，即

$$F(X=x)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x}e^{\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}dX \qquad (17\text{-}2)$$

$F(X)$ 被称为累积分布函数（cumulative distribution function，CDF）。

当 μ=0、σ=1 时，正态分布的概率密度函数（PDF）与累积分布函数（CDF）的关系见图 17-4。公式（17-2）的积分计算比较复杂，但利用 Excel 内置函数计算就相当方便。公式（17-2）对应的 Excel 函数格式为：

NORMDIST（变量值 X，均数 μ，标准差 σ，1）

其中 NORMDIST 函数中的 1 为逻辑值，指明函数的形式。如果为 1，则函数 NORMDIST 返回累积分布函数（CDF），即公式（17-2））$F(X)$，即从 $-\infty$ 开始累积到 X 值的曲线下面积；如果为 0，返回概率密度函数（PDF），即公式（17-1））$f(X)$，即获得正态分布曲线的纵轴坐标值。

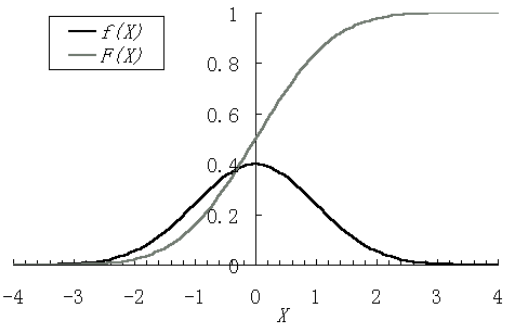


图 17-4　概率密度函数与累积分布函数的关系

对于例 17-2，如果将 X=2400、μ=3250、σ=360，分别存放在单元格 A1、B1、C1 中。单击单元格 E2，键入 Excel 计算公式“=NORMDIST(A1,B1,C1,1)”，按 Enter 键，可获得结果为 0.00911。这就是说，该地当年体重低于 2400 克的婴儿占 0.911%。

例 17-3：某研究者获得某地某年 200 名婴儿出生体重的均数为 3250 克，标准差为 360 克，试根据正态分布原理估计该地当年出生体重在 2800～3500 克的婴儿占多大比例？该体重范围内所含人数是多少？

解：采用 Excel 公式“=NORMDIST(3500,3250,360,1)–NORMDIST(2800,3250,360,1)”，获得 0.6506，即出生体重在 2800～3500 克的婴儿占 65.06%，该体重范围的人数为 65.06%×200≈130 人。

在许多医学研究中，常对一组“正常人”的某项指标进行测定，以确定该指标的参考值范围（reference range）。当某项指标测量值的直方图接近正态分布时，可利用正态分布原理估计“正常人”的某项指标总体的 95%或 99%范围。μ 和 σ 通常未知，只能采用样本均数 $\bar{X}$ 和样本标准差 S 来代替，以（$\bar{X}-1.96S$，$\bar{X}+1.96S$）、（$\bar{X}-2.58S$，$\bar{X}+2.58S$）分别作为该项指标的 95%、99%参考值范围估计值。

17.1.3　正态分布综合电脑实验

这个电脑实验与前面的不同之处在于：

①使用 Excel 函数产生正态分布的概率密度函数（PDF）与累积分布函数（CDF）结果；

②为了比较，将概率密度函数与累积分布函数绘制的图形放在同一界面；

③为了获得动态效果，采用了 Excel 控件工具条（如图 17-5 所示）。

为了获得绘制正态分布曲线的数据，可按下列具体步骤进行操作。

① 按如图 17-5 所示键入单元格区域 A2:C2、E2、F2:F4 中的文本。

② 为了使绘出的曲线两端不至于太长或太短，X 的最小值与最大值分别为（均数±4 标准差），即–4～＋4（均数为 0，标准差为 1）。X 的步长以（标准差/100）为宜，即 0.01。单元格 A3 的值为–4，单元格 A4 的值为–3.99，选取单元格区域 A3:A4，向下拖放填充至单元格 A803。

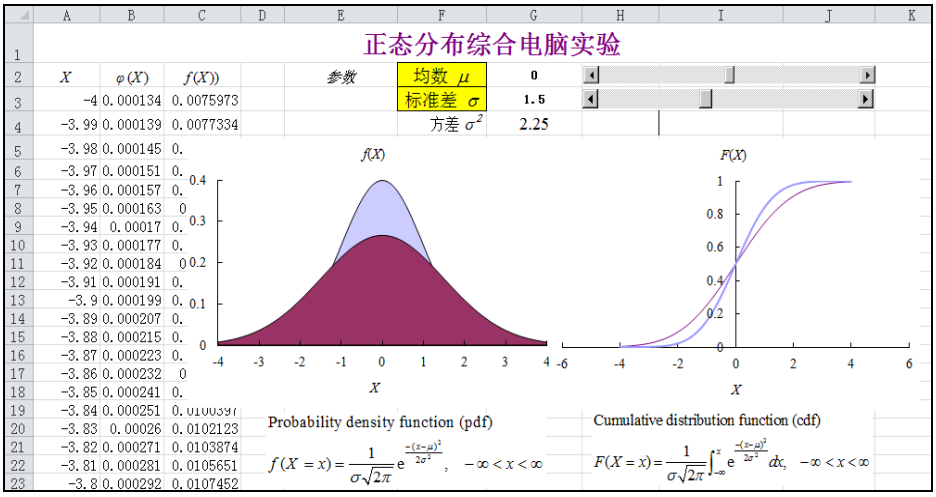


图 17-5　正态分布综合电脑实验

③ 单击单元格 B3，在此单元格（根据公式（17-1））中键入 Excel 公式“=NORMDIST(A3,0,1,0)”，向下拖放填充至单元格 B803，获得标准正态分布曲线的纵坐标值；（根据公式（17-1））在单元格 C3 中键入 Excel 公式“=NORMDIST(A3,$G$2,$G$3,0)”，向下拖放填充至单元格 C83，获得由单元格 G2（均数）、G3（标准差）定义的正态分布曲线纵坐标值。

在单元格 L3、M3 中分别键入 Excel 公式“=NORMDIST(A3,0,1,1)”、“=NORMDIST(A3,$G$2,$G$3,1)”，选取单元格区域 L3:M3，向下拖放填充至单元格 M83，获得制作累积分布函数曲线所需的数据。G2、G3（均、标准差）随着滚动条的滚动发生改变。滚动条由“开发工具→插入→Active X 控件”的第一行的最后一个“滚动条（Active X 控件）”，在需要制作“滚动条”处拉放至适当大小。在“滚动条”上单击鼠标右键，单击“属性”，对“属性”参数进行设置。例如，在均数的滚动条中，修改“单元格链接（LinkedCell）”为 H2，“最大值（Max）”为 60，“最小值（Min）”为 0，步长（LargeChange）默认为 1，在 G2 中键入“(H2–30)/10”，使得改变滚动条后，均数从–3 到+3 变化，步幅为 0.1。

图 17-5 中的概率密度函数曲线采用图表向导中的“面积图”（如图 17-6 所示）绘制而成，其数据在 A、B、C 三列。而累积分布函数曲线仍然采用“*XY* 散点图”绘制，其数据在 A、L、M 三列。

电脑实验

滑动均数滚动条，可见正态分布的概率密度函数 $f(X)$ 曲线左右移动，所以均数是位置参数，反映数据的集中位置；滑动标准差滚动条，可见正态分布的密度函数 $f(X)$ 曲线左右位置不变，但曲线形状改变，标准差越小曲线越瘦高（集中），标准差越大曲线越矮胖（离散），所以标准差是形状参数，反映数据的离散程度。

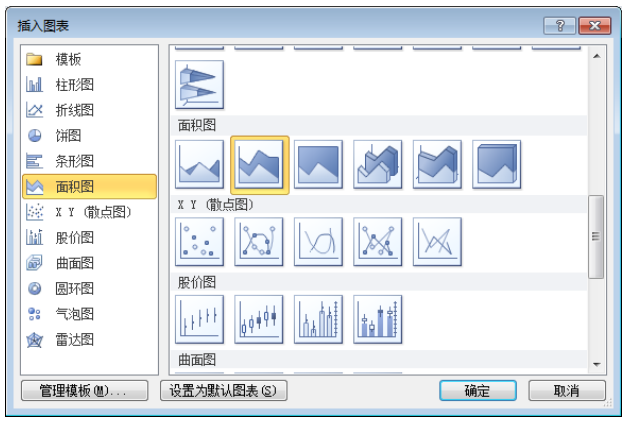


图 17-6　面积图的图表向导

累积分布曲线也有类似改变，即改变均数，曲线位置改变，形状不变；改变标准差，曲线位置不变，形状改变。累积分布函数曲线是概率密度函数曲线下面积从左至右的累积值，是单调上升的曲线。曲线下面积最小值为0，最大值为1。

17.2 二项分布

某一次某诊断试验的结果为阳性或阴性，某一次某病人接受某种治疗的结果为有效或无效，某一次某临床随访观察病人的结果为存活或死亡，某一次某病人对医生的服务评价为满意或不满意，等等。这些只有两个互斥结果的一次性试验称为伯努利（Bernoulli）试验。将伯努利试验重复 n 次，获得“成功”次数为 X 次（“失败”次数为 n-X），常称 X 服从二项分布。二项分布的特点是：

①每项研究感兴趣变量的结果只有两种可能，且相互排斥（A与非A）。

②每一次观察个体的结果为两个互相排斥的结果之一（A或非A）。

③每一观察个体的条件保持不变。即每一观察个体发生A的概率保持不变，均为常数 π。

④每一观察个体间相互独立。即某观察个体将出现什么样的结果与其他个体出现的结果无关，互相不影响。

17.2.1 概率质量函数

令发生事件A（如阳性）的概率为 π，那么发生非A（如阴性）的概率为1-π，n 次试验得到 x 次为A（如阳性）的概率质量函数（probability mass function，PMF）为：

$$f\left(X=x\right)=\binom{n}{x}\pi^{x}(1-\pi)^{n-x},\quad X=0,\ 1,\ \cdots,\ x,\ \cdots,\ n \tag{17-3}$$

式中，$\binom{n}{x}=\frac{n!}{x!(n-x)!}$ 为组合（参见第2章）；x 是变量 X 的取值，为正整数。公式（17-3）就是二项分布的概率函数，并且称相应的随机变量 X 服从二项分布（binomial distribution），记为 $X\sim B$（n，π）。二项分布概率质量函数有两个参数，一个是总体率 π，一个是试验（观察）次数 n。

概率质量函数（PMF）是离散随机变量在各特定取值上的概率。它和概率密度函数（PDF）的不同之处在于：概率密度函数是对连续随机变量定义的，它取某个特定值的概率为0，只有对连续随机变量在某个区间内的取值进行积分后才得到有实际意义的概率。

公式（17-3）的Excel函数格式为：BINOMDIST（“成功”次数 x，试验总次数 n，总体率 π，0）。

17.2.2 累计分布函数

累计分布函数（cumulative distribution function，CDF）有时也称为概率分布函数或分布函数。二项分布的累计分布函数为：

$$F(X=x)=\sum_{i=0}^{x}\binom{n}{i}\pi^{i}(1-\pi)^{n-i}=P(X\leqslant x),\quad i=0,1,\cdots,x \qquad (17\text{-}4)$$

这是变量 X 取小于等于 x 值的概率累加结果。例如在 x=3 情况下，二项分布的累计分布函数所对应结果为 i=0、1、2、3 的概率之和。

离散随机变量对应的累计分布函数曲线同样是不连续的曲线，曲线呈阶梯状。

如果 BINOMDIST 函数中的最后一个选项 0 改变为 1，那么将获得累计分布函数（CDF）$F(X=x)$，即：

BINOMDIST(“成功”例数 x，试验次数 n，概率 π，1)

累计分布函数的逆函数：

CRITBINOM(试验次数 n，概率 π，$P(X\leqslant x)$)

可反过来由概率获得例数 x 的值。

比如由累计分布函数“=BINOMDIST(10,20,0.7,1)”得到累计概率为 0.04796；反过来可由累计分布函数的逆函数“=CRITBINOM(20,0.7,0.04796)”得到 10。

$P(X\geqslant x)$、$P(X<x)$、$P(X>x)$ 也都是累计分布函数，可分别通过 $1-P(X\leqslant x)+P(X=x)$、$P(X\leqslant x)-P(X=x)$、$1-P(X\leqslant x)$ 获得。

17.2.3 二项分布电脑实验

例 17-4：见光盘文件 17-3 两个二项分布分别为 B（50,0.5）和 B（n,p），后者是一个可变的二项分布，试制作电脑实验比较这两个分布，并总结二项分布的规律性。

解：为了获得表达这两个二项分布的图形，首先根据概率函数建立图形数据。按如下步骤建立数据。

① 在单元格 A3、A4 中分别键入 0、1，选取单元格区域 A3:A4，将鼠标移至右下角，出现实心十字时，向下拖放至 A53，获得“成功”次数变量 X；在单元格 B3 中键入公式“=BINOMDIST($A3,50,0.5,0)”，将此公式向下拖放至 B53，获得二项分布 B(50,0.5)对应于“成功”次数变量 X 的概率；在单元格 C3 中键入公式“=BINOMDIST($A3,$G$2,$G$3,0)”，将此公式向下拖放至 C53，获得二项分布 B(n,p)对应于“成功”次数变量 X 的概率。其中试验次数 n 与样本“成功”概率 P 分别由单元格 G2、G3 控制，单元格 G2、G3 又由“开发工具→插入”建立的“滚动条”获得。控制单元格 G2 的滚动条取值为 1～50 正整数，

步幅为 1；控制单元格 G3 的滚动条取值为 0.01～1.00，步幅为 0.01。

② 以上是绘制概率质量函数（PMF）所需的数据。为了获得累计分布函数图，需要像第 6 章绘制生存曲线一样，每一个 X 值需要重复出现两次（见光盘文件“17-3 二项分布.xlsx”的第 N 列，除单元格 N3 外）。在单元格 L3 中键入“=BINOMDIST(0,50,0.5,1)”，在单元格 L4 中键入“=BINOMDIST($N3,50,0.5,1)”，将单元格 L4 公式向下拖放至 L103，建立二项分布 $B(50,0.5)$对应于“成功”次数变量 X 的累计分布概率数据；同样，在单元格 M3 中键入“=BINOMDIST(0,$G$2,$G$3,1)”，在单元格 M4 中键入“=BINOMDIST($N3,50,0.5,1)”，将单元格 M4 公式向下拖放至 M103，建立二项分布 $B(n,p)$ 对应于“成功”次数变量 X 的累计分布概率数据。

采用簇状柱形图（第一个）绘制二项分布概率质量函数（PMF）图形，采用无数据点折线散点图（最后一个）绘制二项分布的累计分布函数（CDF）曲线，绘制的图形如图 17-7 所示。

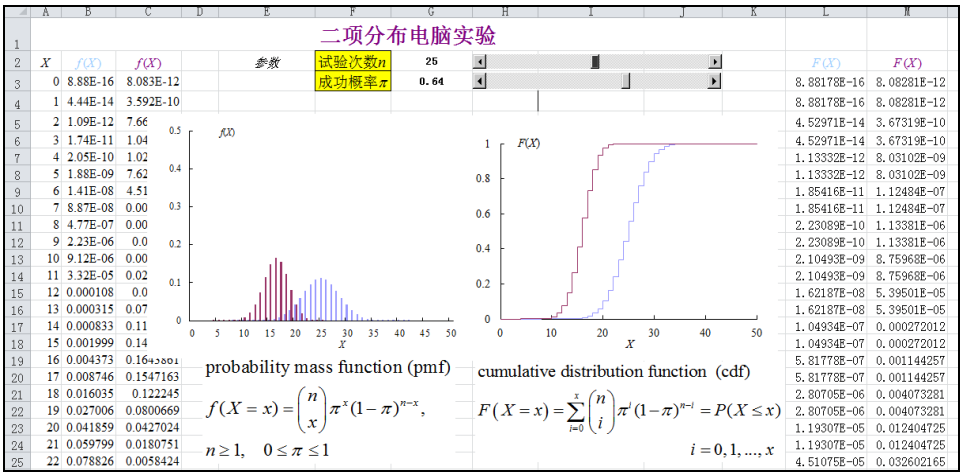


图 17-7　二项分布电脑实验

电脑实验

滑动试验次数 n 滚动条，观察二项分布的概率质量函数 $f(X)$ 图的变化。随着试验次数 n 的增加，图形位置从左向右移动，图形形状从偏态转化为正态。滑动成功概率 π 滚动条，观察二项分布的概率质量函数 $f(X)$ 图，成功概率 π 越小于 0.5，图形越呈右（正）偏态；成功概率 π 越大于 0.5，图形越呈左（负）偏态；成功概率 π=0.5，无论试验次数 n 为多少，均呈对称分布。累计分布曲线的位置改变与概率质量函数 $f(X)$ 图类似。

通过改变二项分布的两个参数观察概率函数图形的改变，可以总结二项分布图形的特

点如下：

①横轴上变量的取值只能是 0，1，2，……等整数；

②图形中间高、两边低；

③如果样本量（试验次数）n 不大，成功概率 π>0.5 时，在横轴的负方向拖一长尾呈负偏峰；当 $\pi<0.5$ 时，在横轴的正方向拖一长尾呈正偏峰；当 π=0.5 时，呈对称分布；

④如果样本量 n 足够大，以至于 $n\pi$ 和 $n(1-\pi)$ 均大于 5，则即使 $\pi\neq0.5$，图形也接近对称。

17.2.4 样本率及率的标准误

若 $X\sim B(n,\pi)$，则 X 的总体均数和总体标准差分别为：

$$\mu_X=n\pi \text{ 和 } \sigma_X=\sqrt{n\pi(1-\pi)}$$

实践中，人们常用样本率 $p=x/n$ 表示二项分布资料。样本率的总体均数和总体标准差分别为：

$$\mu_p=\pi \text{ 和 } \sigma_p=\sqrt{\frac{\pi(1-\pi)}{n}} \qquad (17\text{-}5)$$

样本率的总体标准差也叫率的标准误。在实际应用中，总体率 π 一般是未知的，常用样本率代替，计算率的标准误可以采用以下计算公式：

$$S_p=\sqrt{\frac{p(1-p)}{n}} \qquad (17\text{-}6)$$

例 17-5：某小学随机抽取 200 人作视力检查，其中 50 人近视，问该小学的近视率和率的标准误为多少？

解：已知 $n=200$，$x=50$；有：

$$p=50/200=0.25=25\%,\quad S_p=\sqrt{\frac{0.25(1-0.25)}{200}}=0.0306$$

即该小学的近视率为 25%，率的标准误为 3.06%。

17.3 Poisson 分布

当 n 很大、π 很小时，二项分布 B（n, π）就近似于 Poisson 分布。Poisson 分布由法国数学家 SD Poisson（1781～1840）首先提出而得名。单位时间（平面、空间）内稀有事件的发生数（记为 X），如在放射性物质中单位时间内的脉冲次数、在充分摇匀的瓶中单位容积细菌数、野外单位空间内的某昆虫数、某地总人口中某年内肝癌死亡数等，一般都服从 Poisson 分布。

17.3.1 概率质量函数与累计分布函数

Poisson 分布的概率质量函数（PMF）为：

$$P(X=x)=\frac{e^{-\lambda}\lambda^{x}}{x!},\ X=0,1,\ \cdots,x,\cdots \qquad (17\text{-}7)$$

其中随机变量 X 的取值为整数 x，e=2.718 为自然对数的底。λ 为均数，是 Poisson 分布的唯一参数，因此常将 Poisson 分布记为 $X\sim P(\lambda)$。Excel 函数的格式为：POISSON（变量 X,参数 λ,0）。如果函数中的 0 改变为 1，则求累计分布函数的概率 $P(X\leqslant x)$。

Poisson 分布有以下性质：

①总体均数等于总体方差；

② Poisson 分布具有可加性；

如果 $X_1,X_2,\cdots,X_k$ 相互独立，且它们分别服从以 $\lambda_1,\lambda_2,\cdots,\lambda_k$ 为参数的 Poisson 分布，则 $T=X_1+X_2+\cdots+X_k$ 也服从 Poisson 分布，其参数为 $\lambda_1+\lambda_2+\cdots+\lambda_k$。

③Poisson 分布的正态近似性；

随 λ 的增大，Poisson 分布的对称性越来越好，此时 Poisson 分布近似于正态分布 $N(\lambda,\lambda)$。

④二项分布的 Poisson 分布近似性。

设 $X\sim B(n,\pi)$，则当 $n\to\infty$ 且 $n\pi=\lambda$ 保持不变时，可以证明 X 的极限分布是以 λ 为参数的 Poisson 分布。

17.3.2 Poisson 分布电脑实验

例 17-6：光盘文件 17-4 两个 Poisson 分布分别为 $P(20)$ 和 $P(\lambda)$，后者是一个可变的 Poisson 分布，试制作电脑实验比较这两个分布，并总结 Poisson 分布的规律性。

解：为了获得表达这两个 Poisson 分布的图形，首先根据概率质量函数建立图形数据。按如下步骤建立数据。

① 在单元格 A3、A4 中分别键入 0、1，选取单元格区域 A3:A4，将鼠标移至右下角，出现实心十字时，向下拖放至 A53，获得变量 X；在单元格 B3 中键入公式“=POISSON($A3,20,0)”，将此公式向下拖放至 B53，获得 Poisson 分布 $P(20)$ 对应于变量 X 的概率；在单元格 C3 中键入公式“=POISSON($A3,$G$2,0)”，将此公式向下拖放至 C53，获得 Poisson 分布 $P(\lambda)$ 对应于变量 X 的概率。其中均数 λ 由单元格 G2 控制，单元格 G2 由“开发工具→插入”建立的“滚动条”获得。控制单元格 G2 的滚动条取值为 1～50 正整数，步幅为 1。

② 以上是绘制概率质量函数（PMF）所需的数据。为了获得累计分布函数图，需要像第 6 章绘制生存曲线一样，每一个 X 值需要重复出现两次（见光盘文件“17-4 Poisson 分布.xlsx”的第 N 列，除单元格 N3 外）。在单元格 L3 中键入“=POISSON（0,20,1）”，在单元格 L4 中键入“=POISSON(N3,20,1)”，将单元格 L4 公式向下拖放至 L103，建立 Poisson 分布 P（20）对应于变量 X 的累计分布概率数据；同样，在单元格 M3 中键入“=POISSON(0,$G$2,1)”，在单元格 M4 中键入“=POISSON(N3,$G$2,1)”，将单元格 M4 公式向下拖放至 M103，建立 Poisson 分布 $P(\lambda)$ 对应于变量 X 的累计分布概率数据。

采用簇状柱形图（第一个）绘制 Poisson 分布概率质量函数（PMF）图形，采用无数据点折线散点图（最后一个）绘制 Poisson 分布的累计分布函数（CDF）曲线，绘制的图形如图 17-8 所示。

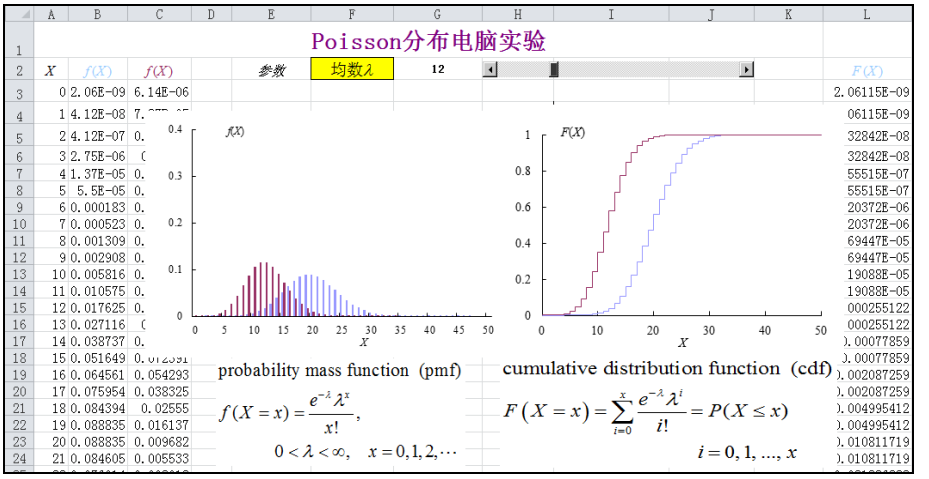


图 17-8　Poisson 分布电脑实验

电脑实验

滑动均数 λ 滚动条，观察 Poisson 分布的概率质量函数 $f(X)$ 图的变化。随着均数 λ 的增加，图形位置从左向右移动，图形形状从偏态转化为正态。Poisson 分布的概率质量函数图也是中间高、两边低，横轴上取值也只能是 0,1,2……整数；当 λ 较小（如不大于 5）时，呈正偏峰；当 λ 较大时，图形逐渐对称。无论何时不可能出现负偏峰图形。

17.4　其他连续分布

除正态分布以外，其他的连续分布有：对数正态分布、均匀分布、指数分布、Weibull

分布、Beta 分布、Gamma 分布等。

17.4.1　对数正态分布

有时也将一些非正态分布变量转化为正态分布数据来处理。若变量明显呈正偏峰分布，如住院费用，可将该变量各观察值直接取对数（如 $\ln X$）；若资料明显呈负偏峰分布，需要通过尝试找出一个合适的常数 k（可用一个略大于最大值的数值），用常数 k 减去各观察值，然后取对数（如 $\ln(k-X)$），若对数值经观察或检验呈正态分布，该分布被称作对数正态分布（logarithmic normal distribution）。对数正态分布的累积分布函数的 Excel 公式格式为“=LOGNORMDIST（X,均数,标准差）”。

光盘文件“17-5 对数正态分布.xlsx”可进行对数正态分布电脑实验，制作方法类似正态分布。分别滑动均数与标准差滚动条，可观察概率密度函数（PDF）曲线和累积分布函数（CDF）曲线的变化（如图 17-9 所示）。图中淡颜色图形固定，且均数 μ=0，标准差 σ=1；深颜色图形为可变对数正态分布。

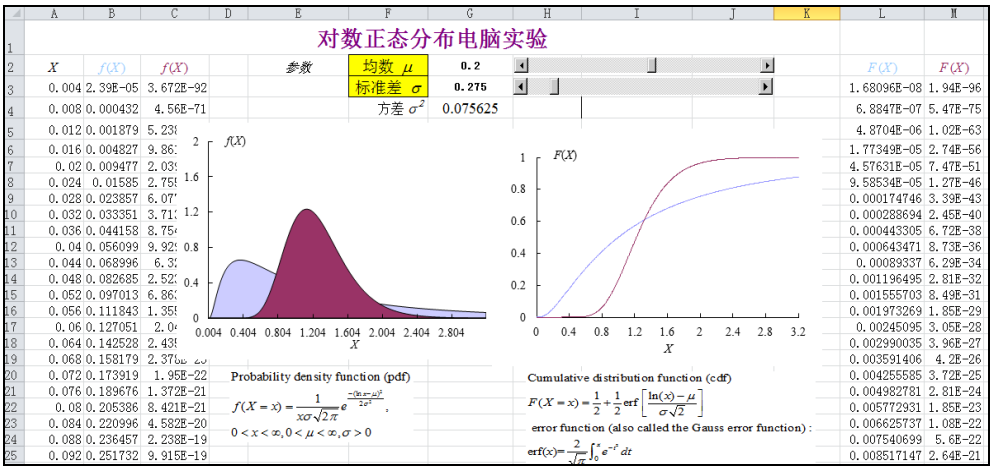


图 17-9　对数正态分布电脑实验

17.4.2　均匀分布

均匀分布（uniform distribution）由下限值 a 与上限值 b 两个参数决定，从 a 到 b 的概率密度函数曲线纵高恒等于 $1/(b-a)$。累积分布曲线呈阶梯状，从 a 到 b 的累积分布曲线纵高等于 $(x-a)/(b-a)$（如图 17-10 所示）。

光盘文件“17-6 均匀分布.xlsx”可进行均匀分布电脑实验。分别滑动下限值 a 与上限值 b 滚动条，可观察概率密度函数（PDF）曲线和累积分布函数（CDF）曲线的变化。图

17-10 中淡颜色图形固定，且下限值 $a=-1$，上限值 $b=1$；深颜色图形为可变均匀分布。

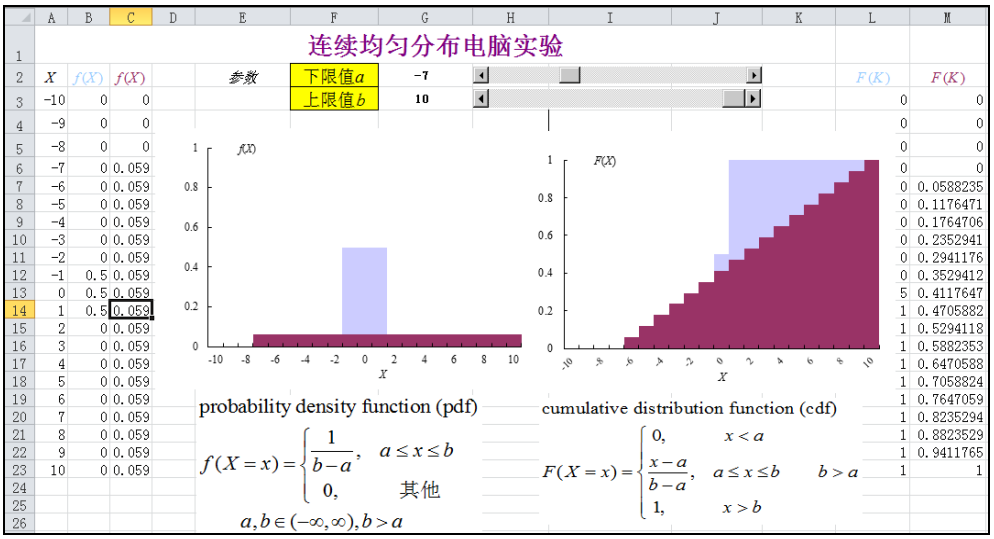


图 17-10　均匀分布电脑实验

17.4.3　指数分布

指数分布只有一个参数，即率 λ。其概率密度函数（PDF）为：

$$f(X=x)=\lambda e^{-\lambda x},\quad 0\leqslant x<\infty,\quad \lambda>0 \tag{17-8}$$

相应的 Excel 公式格式为“=EXPONDIST（X,率,0）”。累积分布函数（CDF）为

$$F(X=x)=1-e^{-\lambda x},\quad 0\leqslant x<\infty,\quad \lambda>0 \tag{17-9}$$

相应的 Excel 公式格式为“=EXPONDIST（X,率,1）”。

光盘文件“17-7 指数分布.xlsx”可进行指数分布电脑实验。滑动参数率 λ 滚动条，可观察概率密度函数（PDF）曲线和累积分布函数（CDF）曲线的变化（如图 17-11 所示）。图中淡颜色图形固定，且参数 $\lambda=1$，深颜色图形为可变指数分布。

17.4.4　Weibull 分布

Weibull 分布有两个参数，即形状参数 k 和尺度参数 λ。其概率密度函数（PDF）为

$$f(X=x)=\frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1}e^{-(x/\lambda)^k},\quad 0\leqslant x<\infty,\quad k>0,\lambda>0 \tag{17-10}$$

相应的 Excel 公式格式为“=WEIBULL（X,形状参数 k,尺度参数 λ,0）”。累积分布函数（CDF）为：

$$F(X=x)=1-\mathrm{e}^{-(x/\lambda)^k},\quad 0\leqslant x<\infty,\quad k>0,\lambda>0 \qquad (17\text{-}11)$$

相应的 Excel 公式格式为“=WEIBULL（X,形状参数 k,尺度参数 λ,1）”。

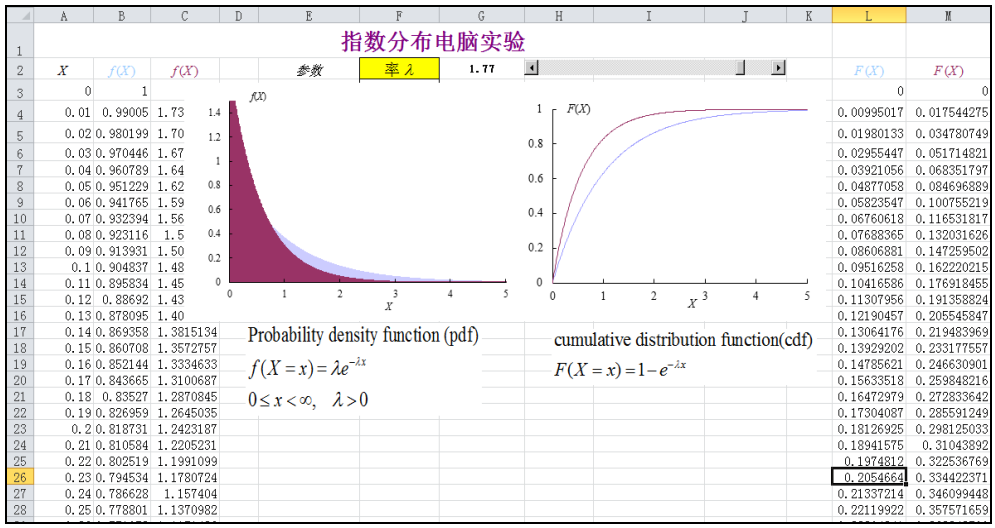


图 17-11　指数分布电脑实验

光盘文件“17-8 weibull 分布.xlsx”可进行指数分布电脑实验。分别滑动形状参数 k 和尺度参数 λ 滚动条，可观察概率密度函数（PDF）曲线和累积分布函数（CDF）曲线的变化（如图 17-12 所示）。图中淡颜色图形固定，且形状参数 k=5，尺度参数 λ=1.5，深颜色图形为可变 Weibull 分布。

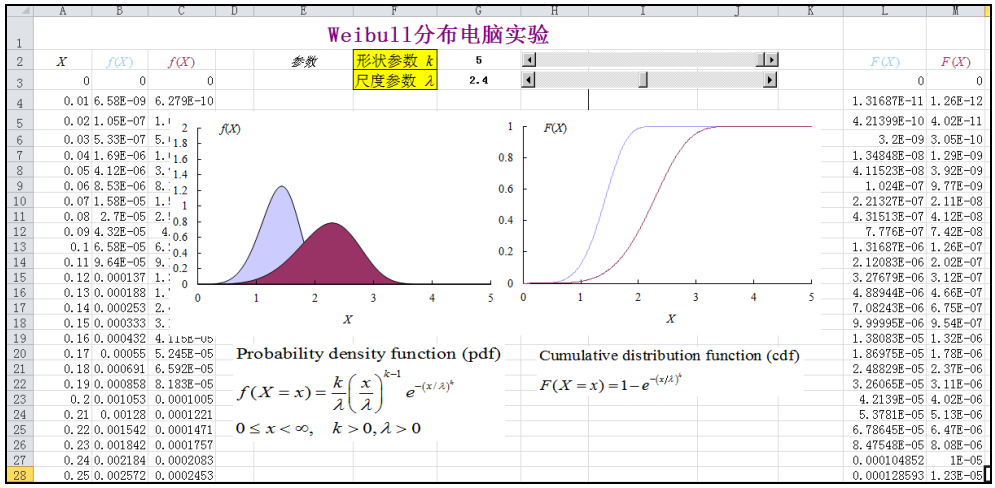


图 17-12　Weibull 分布电脑实验

17.4.5 Beta 分布

Beta 分布有两个参数，即形状参数 α 和尺度参数 β。其概率密度函数（PDF）为：

$$f(X=x)=\frac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}$$
$$=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1},\quad 0\leqslant x\leqslant 1,\quad \alpha>0,\beta>0 \tag{17-12}$$

累积分布函数（CDF）为：

$$F(X=x)=\frac{B_x(\alpha,\beta)}{B(\alpha,\beta)} \tag{17-13}$$

相应的 Excel 公式格式为“=BETADIST（X,形状参数 α ,尺度参数 β ）”。

光盘文件“17-9 Beta 分布.xlsx”可进行 Beta 分布电脑实验。分别滑动形状参数 α 和尺度参数 β 滚动条，可观察概率密度函数（PDF）曲线和累积分布函数（CDF）曲线的变化（如图 17-13 所示）。图中淡颜色图形固定，且形状参数 α =5，尺度参数 β =5；深颜色图形为可变 Beta 分布。

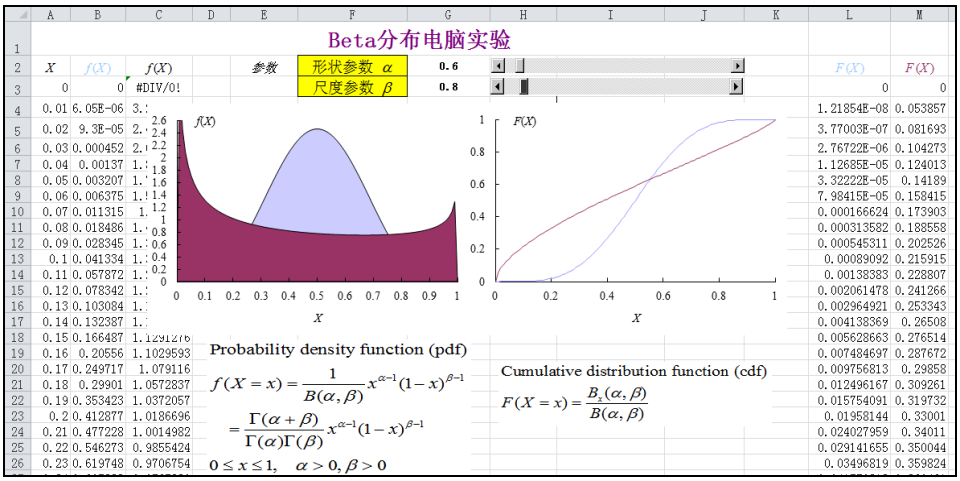


图 17-13　Beta 分布电脑实验

17.4.6 Gamma 分布

Gamma 分布有两个参数，即形状参数 k 和尺度参数 θ。其概率密度函数（PDF）为：

$$f(X=x)=x^{k-1}\frac{e^{-x/\theta}}{\Gamma(k)\theta^{k}},\quad 0\leqslant x<\infty,\quad k>0,\theta>0 \tag{17-14}$$

相应的 Excel 公式格式为“=GAMMADIST（X,形状参数 k,尺度参数 θ ,0）”。累积分布函数（CDF）为：

$$F(X=x)=\frac{\gamma(k,x/\theta)}{\Gamma(k)},\quad 0\leqslant x<\infty,\quad k>0,\theta>0 \tag{17-15}$$

相应的 Excel 公式格式为“=GAMMADIST（X,形状参数 k,尺度参数 θ ,1）”。

光盘文件“17-10 Gamma 分布.xlsx”可进行 Gamma 分布电脑实验。分别滑动形状参数 k 和尺度参数 θ 滚动条，可观察概率密度函数（PDF）曲线和累积分布函数（CDF）曲线的变化（如图 17-14 所示）。图中淡颜色图形固定，且形状参数 k=5，尺度参数 θ =1；深颜色图形为可变 Gamma 分布。

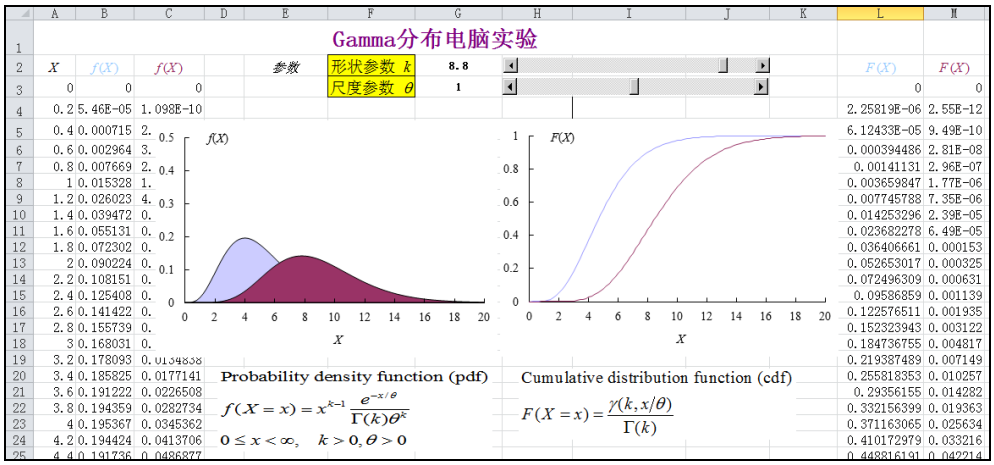


图 17-14　Gamma 分布电脑实验

17.5　其他离散分布

除二项分布、Poisson 分布以外，其他的离散分布有：均匀分布、超几何分布等。

17.5.1　均匀分布

在 17.4 节“其他连续分布”中已经阐述了均匀分布。此外，均匀分布也可以是离散的，其变量 X 的取值可以是不连续的。离散均匀分布的概率质量函数（PMF）为：

$$f(X=x)=\begin{cases}\dfrac{1}{b-a+1}, & a\leqslant x\leqslant b\\ 0, & \text{其他}\end{cases}\quad a,b\in(\cdots,-2,-1,0,1,2\cdots),\ x\in(a,a+1,\cdots,b-1,b) \tag{17-16}$$

累计分布函数（CDF）为：

$$F(X=x)=\begin{cases}0, & x<a\\ \dfrac{x-a+1}{b-a+1}, & a\leqslant x\leqslant b\\ 1, & x>b\end{cases} \qquad (17\text{-}17)$$

光盘文件“17-11 均匀分布 ls.xlsx”可进行离散均匀分布电脑实验。分别滑动下限值 a 与上限值 b 滚动条，可观察概率质量函数（PMF）曲线和累计分布函数（CDF）曲线的变化。图 17-15 中淡颜色图形固定，且下限值 a=0，上限值 b=1；深颜色图形为可变均匀分布。

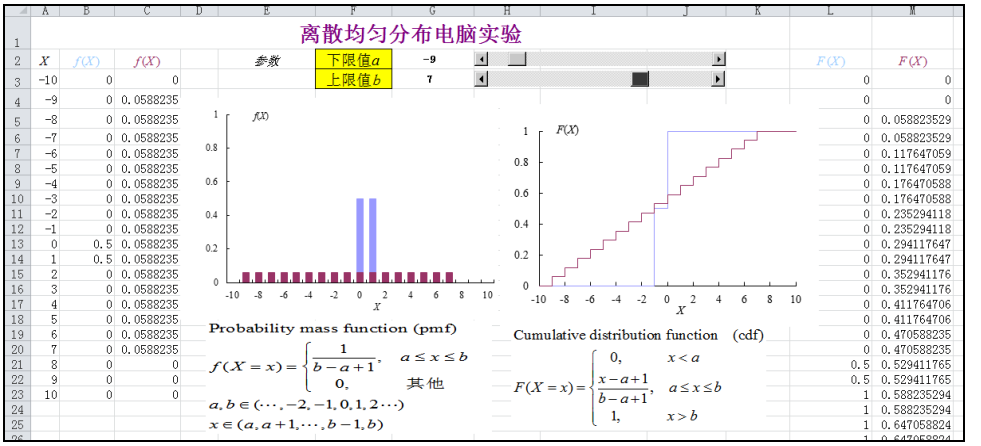


图 17-15　离散均匀分布电脑实验

17.5.2　超几何分布

超几何分布（hypergeometric distribution）是进行 Fisher 精确检验的基础，描述了在有限总体（N 个个体中含 m 个“成功”）中，无放回抽取 n 个个体，其中有 x 个“成功”的概率。超几何分布的概率质量函数（PMF）为：

$$f(X=x)=\frac{\binom{m}{x}\binom{N-m}{n-x}}{\binom{N}{n}}, \qquad \begin{array}{l}x\in\max(0,n+m-N),\cdots,\min(m,n)\\ N\in 0,1,2,\cdots;\quad m,n\in 0,1,2,\cdots,N\end{array} \qquad (17\text{-}18)$$

相应的 Excel 公式格式为“=HYPGEOMDIST（X,样本含量 n,总体成功数 m,总体例数 N）”。累计分布函数（CDF）为：

$$F(X=x)=\sum_{i=0}^{x}f(X=i) \qquad (17\text{-}19)$$

光盘文件“17-12 超几何分布.xlsx”可进行超几何分布电脑实验。分别滑动样本含量 n、

总体成功数 m、总体例数 N 滚动条，可观察概率质量函数（PMF）曲线和累计分布函数（CDF）曲线的变化。图 17-16 中淡颜色图形固定，且样本含量 n=25，总体成功数 m=25，总体例数 N=100；深颜色图形为可变超几何分布。

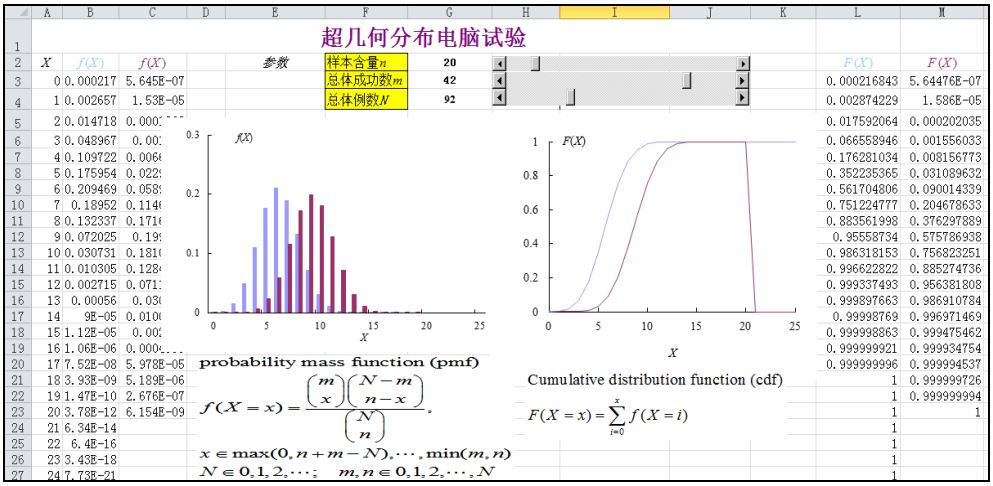


图 17-16　超几何分布电脑实验

第 18 章

随机抽样与抽样误差

本章首先介绍产生随机数以及利用随机数抽取样本的方法，然后介绍随机抽样电脑实验，帮助读者理解与抽样误差相关的统计学概念。

18.1　随机数发生器与随机函数

在 Excel 中，可用“随机数发生器”和随机数函数产生随机数。

18.1.1　随机数发生器

单击“数据→分析→数据分析”（参见 7.1 节），可在分析工具中找到“随机数发生器”（如图 18-1 所示），选择此选项，单击“确定”按钮，将弹出“随机数发生器”对话框（如图 18-2 所示）。

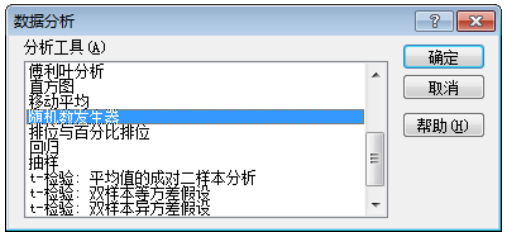


图 18-1　调用随机数发生器

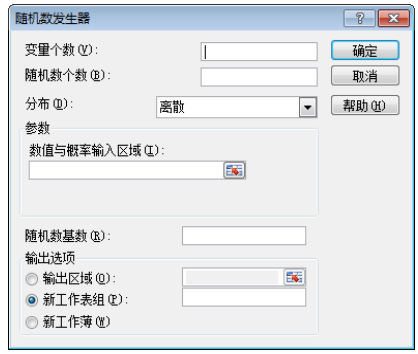


图 18-2　“随机数发生器”对话框

在“随机数发生器”对话框中，在“变量个数”的右侧指定拟输出随机数的列数；在“随机数个数”的右侧指定每一列中拟产生随机数的个数。

Excel 2010 有 7 个产生“随机数”的概率分布类型，分别为正态分布、二项（二项式）分布、Poisson（泊松）分布、均匀分布、Bernoulli（伯努利）分布、模式分布和离散分布（括号内为 Excel 所给名称，以下同）。

1. 正态分布

正态分布有均数、标准差两个参数，默认情况下 Excel 产生均数、标准差分别为 0、1 的标准正态分布随机数。在没有设定随机数基数的情况下，每次产生不同的随机数；随机数基数一旦固定，每次将产生相同的随机数。

如果变量个数为 2，随机数个数为 20，分布选择“正态”，输出区域分别选择 A1、D1，单击“确定”按钮，两次产生的随机数如图 18-3 所示。

如果变量个数为 2，随机数个数为 20，分布选择“正态”，输出区域分别选择 A1、D1，随机数基数选择 11899，单击“确定”按钮，两次产生的随机数如图 18-4 所示。由此可见，随机数基数一旦固定，所产生的随机数也将保持不变。随机数基数可以是任何正整数，以

后重新使用该基数，可生成相同的随机数，便于他人重复后得出相同数据。改变正态分布的均数（平均值）和标准差（标准偏差），可产生不同正态分布的随机数。

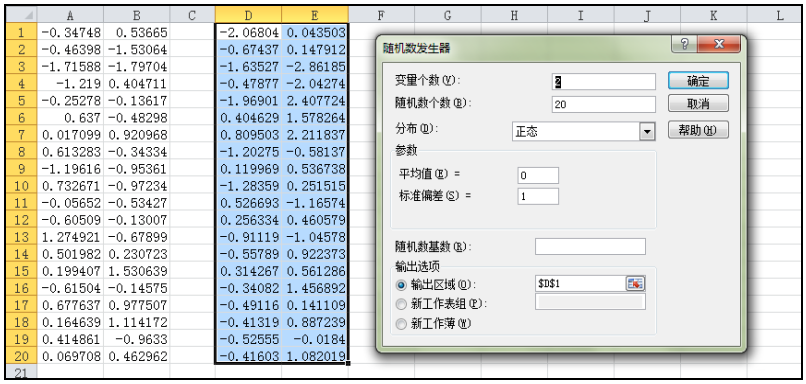


图 18-3　无随机数基数设定情况下产生的“正态”分布随机数

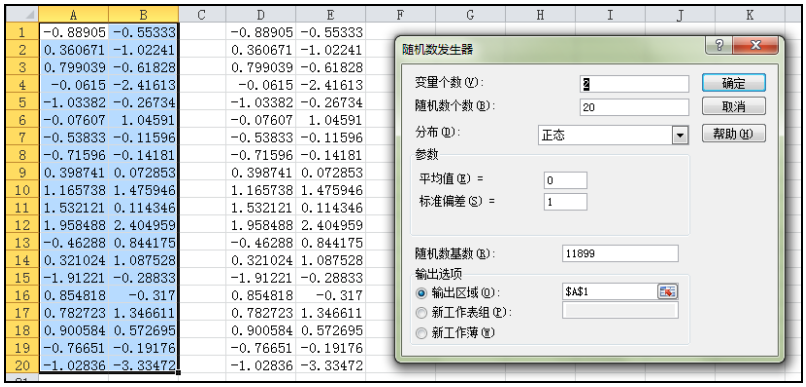


图 18-4　随机数基数固定后的“正态”分布随机数

2. 二项分布

二项分布有试验次数 n、总体率 π 两个参数，产生 0～n 范围内的正整数随机变量。如果“变量个数”为 2，“随机数个数”为 20，“分布”选择“二项式”，总体率 π（p(A)）=0.3，“试验次数(N)”=100，“随机数基数”输入 123，“输出区域”选 A1，单击“确定”按钮，产生的随机数如图 18-5 所示。如果“随机数基数”为空白情况下，每次产生的随机数是变化的；一旦随机数基数固定，每次产生的随机数相同。改变二项分布的参数总体率 π（p(A)）和试验次数(N)，可产生不同二项分布的随机数。

3. Poisson 分布

Poisson 分布只有一个参数 λ，它既是总体均数，也是总体方差（Poisson 分布的均数

与方差相等），产生出的随机变量 X 取值范围为 0、1、2、……。如果“变量个数”为 2，“随机数个数”为 20，“分布”选择“泊松”，“参数 λ(L)”=5，“随机数基数”输入 123，“输出区域”选 A1，单击“确定”按钮，产生的随机数如图 18-6 所示。在随机数基数空白情况下，每次产生的随机数是变化的；一旦随机数基数固定，每次产生的随机数相同。改变 Poisson 分布的参数 λ，可产生不同 Poisson 分布的随机数。

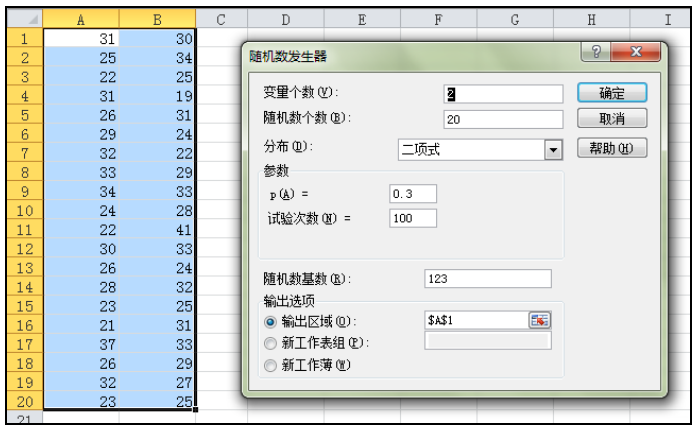


图 18-5　二项分布随机数的产生

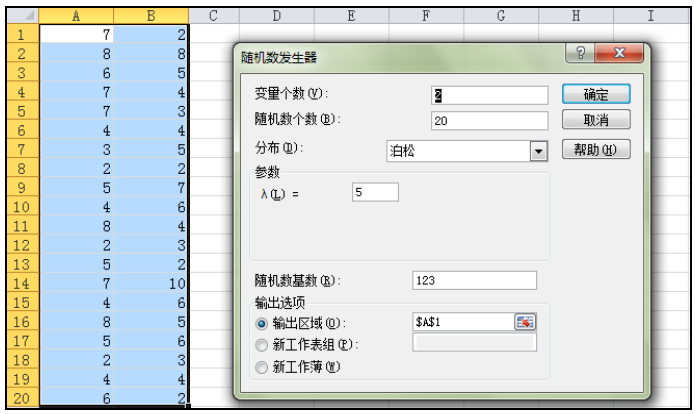


图 18-6　Poisson 分布随机数的产生

4. 均匀分布

均匀分布产生出的随机数将均匀分布于指定的下限值与上限值之间，这就是说，在下限值与上限值之间的所有实数均有相同的概率被抽取（如图 18-7 所示）。如果指定下限值为 0、上限值为 1，那么所产生随机数的功能将与 Excel 随机函数“RAND()”等价。

如果“变量个数”为 2，“随机数个数”为 20，“分布”选择“均匀”，“参数”介于 1

（下限值）与 10（上限值）之间，“随机数基数”输入 123，“输出区域”选 A1，单击“确定”按钮，产生的随机数如图 18-8 所示。在随机数基数空白情况下，每次产生的随机数是变化的；一旦随机数基数固定，每次产生的随机数相同。改变均匀分布的下限值与上限值，可产生不同均匀分布的随机数。

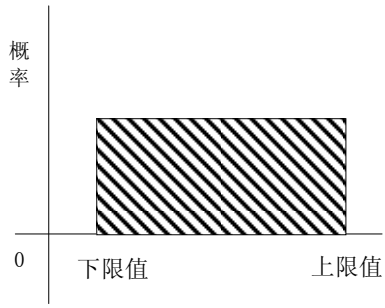


图 18-7　均匀分布示意图

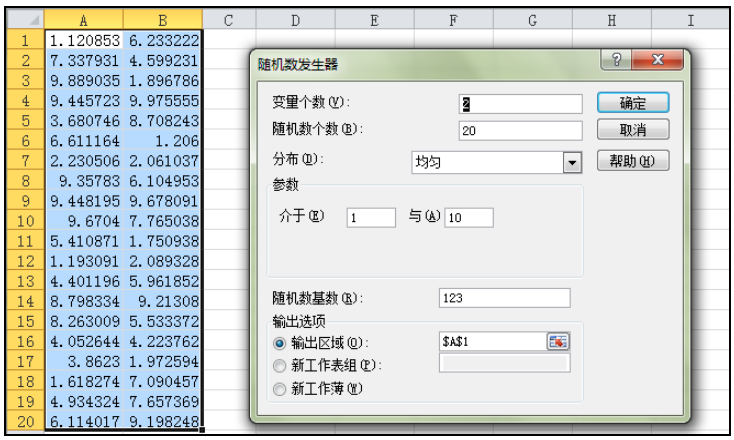


图 18-8　均匀分布随机数的产生

5. Bernoulli 分布

Bernoulli（伯努利）分布，也叫两点分布（即 0、1 两点），其试验叫做 Bernoulli 试验，重复进行 n 次 Bernoulli 试验，出现“1”的次数服从二项分布（参见第 17 章）。Bernoulli 分布产生的随机数只有 0 与 1 两个取值。其参数只有 1 个，即总体率 π。相当于在 0～1 范围的值中均匀产生随机数，如果变量小于或等于总体率 π，则 Bernoulli 随机变量的值为 0；否则，随机变量的值为 1。

如果“变量个数”为 2，“随机数个数”为 20，“分布”选择“伯努利”，指定总体率 π（p(A)）=0.3，“随机数基数”输入 123，“输出区域”选 A1，单击“确定”按钮，产生的随机数

如图 18-9 所示。在随机数基数空白情况下，每次产生的随机数是变化的；一旦随机数基数固定，每次产生的随机数相同。改变 Bernoulli 分布的总体率 π（p(A)），可产生不同 Bernoulli 分布的随机数。

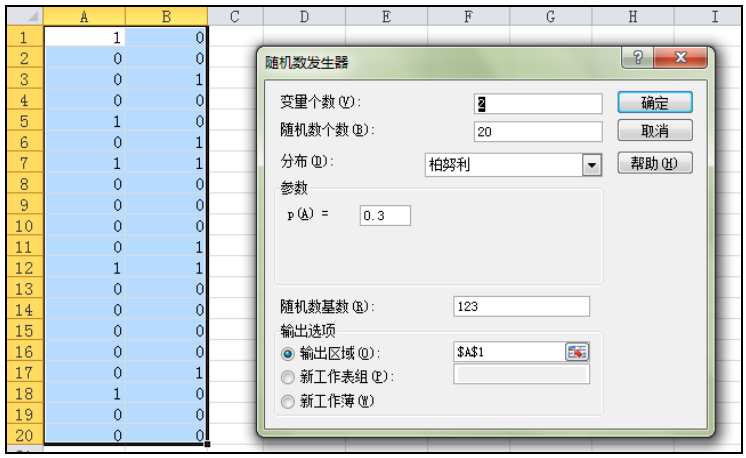


图 18-9 Bernoulli 分布随机数的产生

以上 5 个分布是 Excel 产生真正随机数的分布，所以在填写分布参数的下方，有一个随机数基数（random seed），而剩下的模式分布和离散分布这两个分布无此选项。

离散分布的参数为“数值与概率输入区域”。为了获得这种分布产生的数据，必须事先在 Excel 工作表的某单元格区域输入两列数据，左边的一列为待生成的数值（如 21、22、23），右边的一列为相应的发生概率，其发生概率的取值在 0～1 之间，且该列的概率之和为 1（如 0.2+0.2+0.6=1），否则 Excel 将出现错误提示信息。如果“变量个数”为 2，“随机数个数”为 20，“分布”选择“离散”，指定上述说明的两列数据，“输出区域”选 D1，单击“确定”按钮，产生的随机数如图 18-10 所示。

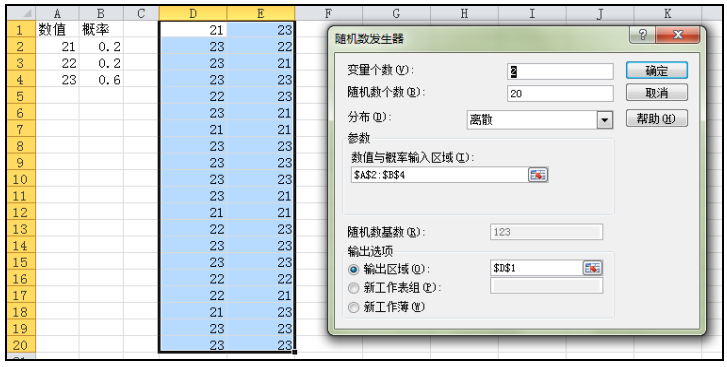


图 18-10 “离散”随机数的产生

以上 6 种分布都是按照某种概率分布产生的数据，而模式分布是按特定设置产生数据的。

模式分布的参数有上界、下界、步长、数值重复次数以及序列重复次数。一旦确定了这些参数，Excel 将按其设置产生数据。在输出选项中，可以具体指定输出区域的起始单元格、工作表等。

18.1.2 随机数函数

Excel 生成随机数的函数为 RAND()或 RANDBETWEEN（下限值,上限值）。这两个函数比分析工具中的“随机数发生器”操作更简单，每按一次 F9 键，随机数将相应发生一次改变。

1. 均匀分布

采用 Excel 公式格式“=下限值+RAND()*（上限值-下限值）”可产生取值在下限值与上限值之间的均匀分布随机数，函数 RAND()实际上是下限值与上限值分别为 0 与 1 的均匀分布随机函数。下限值与上限值需要人为给定，例如，下限值与上限值分别为 1 与 5，其产生均匀分布随机数的 Excel 公式为“=1+RAND()*(5−1)”。

公式“=下限值+RAND()*（上限值-下限值）”与“=RANDBETWEEN（下限值,上限值）”产生随机数的作用基本相同，不同的是 RANDBETWEEN 函数产生整数，而 RAND 函数产生小数。

将以上公式向上、下、左、右拖放填充均可，用鼠标放在工作表的任意位置，每按一次 F9 键，随机数将改变一次。下限值与上限值也可以是单元格引用。这些性质对于以下所给的分布均拥有，所以下面不再提及。

2. 正态分布

采用 Excel 公式格式“=NORMINV（RAND(),均数,标准差）”可产生正态分布随机数。例如，均数与标准差分别为 4.79 与 0.46，其产生正态分布随机数的 Excel 公式为“=NORMINV(RAND(), 4.79, 0.46)”。

3. 二项分布

采用 Excel 公式格式“=CRITBINOM（试验次数 n,概率 π,RAND()）”可产生二项分布随机数，例如，试验次数 n 与概率 π 分别为 20 与 0.7，其产生二项分布随机数的 Excel 公式为“=CRITBINOM(20, 0.7, RAND())”。

4. Bernoulli 分布

采用 Excel 公式格式“=CRITBINOM(1,概率 π,RAND())”可产生 Bernoulli 分布随机数。例如，概率 π 为 0.7，其产生 Bernoulli 分布随机数的 Excel 公式为“=CRITBINOM(1, 0.7,

RAND())”。

对于随机数发生器中的其他三种分布（Poisson、模式、离散分布），采用 RAND() 函数操作难以实现。

18.2 有放回抽样与无归还抽样

抽样大致可分成有放回抽样与无归还抽样两大类。

18.2.1 “抽样”分析工具

对于已知有限总体，从中抽取数据可单击“数据→分析→数据分析→抽样”命令。这种方式的随机抽样为“有放回抽样”（Bootstrap 抽样方法就是这样一种抽样）。

例 18-1：某教研室共有 15 名教师，请从中随机抽取 5 名参加某项活动。

解：首先，将这 15 名教师编号，依次编为第 1、2、……、15 号，分别放在单元格区域 A2:A16；其次，分别在单元格 A1、C1 处输入“教师编号”、“抽样”；然后单击“数据→分析→数据分析→分析工具：抽样”命令，将弹出一个“抽样”对话框，在对话框中选取输入区域 A1:A16，“随机样本数”框内填写 5，“输出区域”框内选取单元格 C2（如图 18-11 所示）。单击“确定”按钮后，在单元格区域 C2:C6 得到的教师编号分别为：3、8、8、8、6。

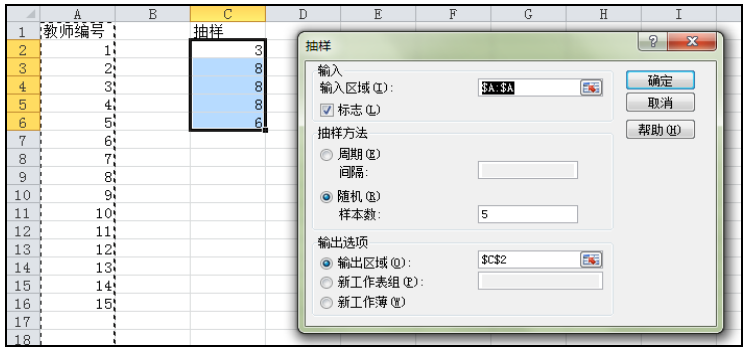


图 18-11 “抽样”对话框

注意事项与结果分析：

①由于抽样是随机的，不同的研究者或同一个研究者不同时间抽取的结果可能不同。

②由于抽样是“有放回抽样”，抽取的样本每次均需归还，第一次抽到后，第二次仍有同等的机会被抽取。因此，在这种抽样获得的样本中，一名教师可能被多次抽取，如本例的 8 号教师仅 5 次抽样就被 3 次抽取。

18.2.2 无归还抽样

例 18-1 本打算抽取 5 名教师，第 8 号教师被 3 次抽取，结果抽取 5 名教师的目的没有达到。那么如何才能使抽取到的个体不再被抽取呢？可采用随机数函数来解决此问题。具体操作步骤如下：

①将这 15 名教师编号，放在单元格区域 A2:A16 中。事先规定：随机数较小的 5 名教师为被选取对象。

②分别在单元格 A1、B1 处输入“教师编号”、“随机数”。

③在单元格 B2 中输入公式“=RAND()”，将鼠标移至该单元格右下角的填充柄处，向下拖放至单元格 B16。

④为了避免随机数函数的“随机”变化，需要对单元格区域 B2:B16 产生的随机数进行复制，然后采用选择性粘贴“数值”到 B2:B16 中。

⑤鼠标放在随机数所在列的任意位置，单击“开始→排序与筛选→升序”命令，Excel 会自动按“随机数”排序。也可在“排序与筛选”中选择“自定义排序”，可对排序依据、排序次序、“数据包含标题”与否等进行设定。按“自定义排序”后，排在前面的 5 位教师的编号，就是我们需要选择的对象。本例选择的教师分别是第 14、3、4、15、9 号（如图 18-12 所示）。

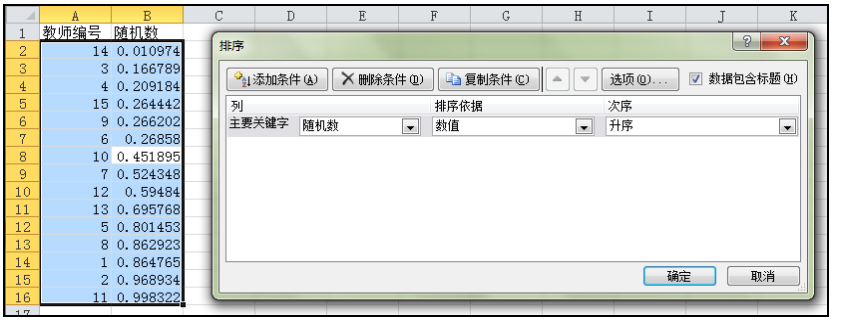


图 18-12 15 名教师的随机选取

对于第⑤步也可采用更简单的操作，先用鼠标单击单元格 B1，即单击选择“随机数”字段名，单击“数据→ AZ↓ ”命令，即可完成排序。

18.3 随机抽样电脑实验

本节将介绍从正态分布或非正态均匀分布总体中，随机抽取样本进行电脑实验，以便理解抽样误差的含义。

18.3.1 从正态总体中随机抽样

1. 产生正态分布随机数据

例 18-2：见光盘文见 18-1 假定正常女子的总胆固醇（mmol/L）服从正态分布 $N(3.73,\ 0.64^2)$，从这个总体中随机抽取 1000 份样本，每份含 n=2 个个体。求每份样本的均数和标准差，并对求得的 1000 个均数绘制直方图。

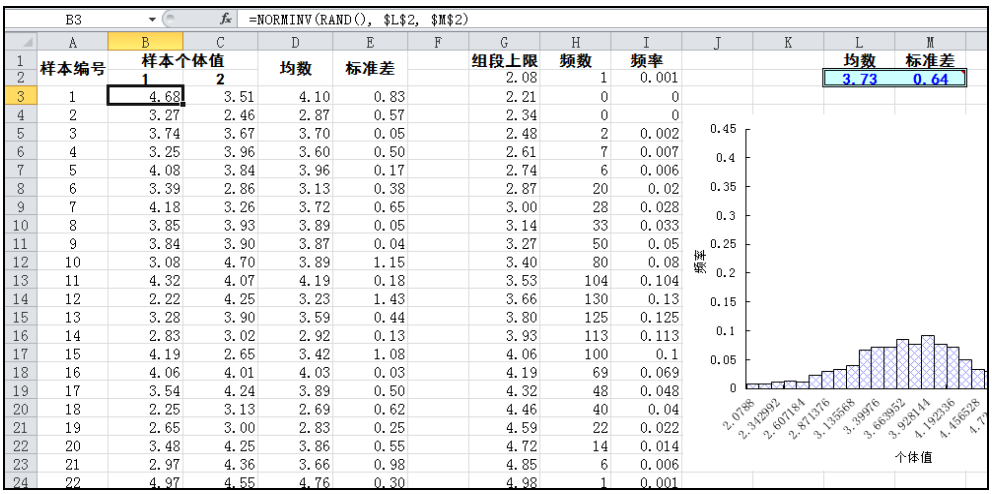
解：按如下步骤产生随机样本，并求均数与标准差。

①在单元格区域 A1:M1 中按图 18-13 分别输入并编辑文本。其中单元格 L2、M2 分别为正态分布的总体均数 μ、总体标准差 σ。

②单击单元格 A3，并输入 1，再单击“开始→编辑→填充→系列”，在弹出的“序列”对话框中，“序列产生在”下选择“列”，并在“终止值”右侧框内输入 1000，步长值默认为 1，单击“确定”按钮，即可在单元格区域 A3:A1002 中产生 1～1000 的序列数。

③在单元格 B3 中输入公式“=NORMINV(RAND(), $L$2, $M$2)”，向右拖放填充该公式至单元格 C3。由此产生第一份样本。

④由第一份样本（B3:C3）计算样本均数与标准差。在单元格 D3 中键入公式“=AVERAGE(B3:C3)”求均数；在单元格 E3 中键入公式“=STDEV(B3:C3)”求标准差。

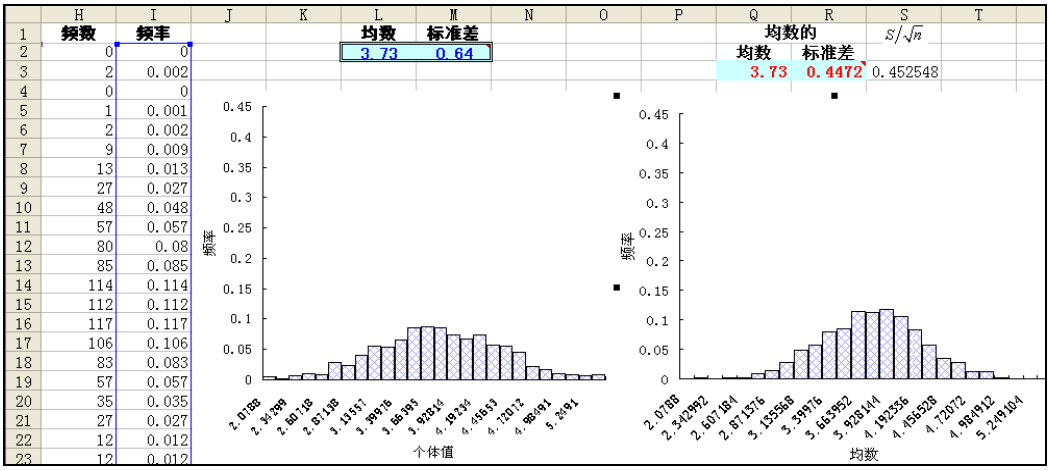


样本编号	样本个体值 1	样本个体值 2	均数	标准差	组段上限	频数	频率	均数	标准差
					2.08	1	0.001	3.73	0.64
1	4.68	3.51	4.10	0.83	2.21	0	0		
2	3.27	2.46	2.87	0.57	2.34	0	0		
3	3.74	3.67	3.70	0.05	2.48	2	0.002		
4	3.25	3.96	3.60	0.50	2.61	7	0.007		
5	4.08	3.84	3.96	0.17	2.74	6	0.006		
6	3.39	2.86	3.13	0.38	2.87	20	0.02		
7	4.18	3.26	3.72	0.65	3.00	28	0.028		
8	3.85	3.93	3.89	0.05	3.14	33	0.033		
9	3.84	3.90	3.87	0.04	3.27	50	0.05		
10	3.08	4.70	3.89	1.15	3.40	80	0.08		
11	4.32	4.07	4.19	0.18	3.53	104	0.104		
12	2.22	4.25	3.23	1.43	3.66	130	0.13		
13	3.28	3.90	3.59	0.44	3.80	125	0.125		
14	2.83	3.02	2.92	0.13	3.93	113	0.113		
15	4.19	2.65	3.42	1.08	4.06	100	0.1		
16	4.06	4.01	4.03	0.03	4.19	69	0.069		
17	3.54	4.24	3.89	0.50	4.32	48	0.048		
18	2.25	3.13	2.69	0.62	4.46	40	0.04		
19	2.65	3.00	2.83	0.25	4.59	22	0.022		
20	3.48	4.25	3.86	0.55	4.72	14	0.014		
21	2.97	4.36	3.66	0.98	4.85	6	0.006		
22	4.97	4.55	4.76	0.30	4.98	1	0.001		

图 18-13 正态分布随机样本及每份样本的均数与标准差

⑤选取单元格区域 B3:E3，在被选取的单元格上单击鼠标右键，在弹出的快捷菜单中选择“复制”命令。选取单元格区域 B4:E1002（单击单元格 B4，用屏幕右侧的滚动条向下滚动直至看到单元格 E1002，按住 Shift 键，再单击单元格 E1002）。在常用工具栏中单

击“粘贴”图标，便可产生 1000 份随机样本，并获得了每份样本（2 个个体值）的均数、标准差。



	H	I	L	M	Q	R	S
1	频数	频率	均数	标准差	均数的		$S/\sqrt{n}$
2	0	0	3.73	0.64	均数	标准差	
3	2	0.002			3.73	0.4472	0.452548
4	0	0					
5	1	0.001					
6	2	0.002					
7	9	0.009					
8	13	0.013					
9	27	0.027					
10	48	0.048					
11	57	0.057					
12	80	0.08					
13	85	0.085					
14	114	0.114					
15	112	0.112					
16	117	0.117					
17	106	0.106					
18	83	0.083					
19	57	0.057					
20	35	0.035					
21	27	0.027					
22	12	0.012					
23	12	0.012					

图 18-14　修饰后的样本均数直方图

⑥在单元格区域 Q1:R2 中，按图 18-14 输入文字“均数的均数、标准差”。在单元格 Q3 中输入公式“=AVERAGE(D3:D1002)”，求 1000 份样本均数的均数；在单元格 R3 中输入公式“=STDEV(D3:D1002)”，求 1000 份样本均数的标准差。

只要光标在产生随机数的工作表中，按一次 F9 键，图 18-13、图 18-14 显示的随机样本及其每份样本的均数标准差（即单元格区域 B3:E1002）、样本均数的均数与其标准差（单元格 Q3、R3）都会发生“随机”的改变。

该实验由此可以得出以下结论：

①样本均数（或标准差）未必等于总体均数（或标准差）；

②样本均数（或标准差）之间存在着差异；

③样本均数的均数在总体均数的上下波动，并且波动幅度比一般样本内个体值之间的波动（本例每份样本 2 个个体）小得多；

④每次按 F9 键，可见样本均数的标准差（即标准误）在 $\sigma/\sqrt{n}=0.64/\sqrt{2}=0.4525$ 上下波动，且波动范围较一般样本标准差波动范围小。

⑤比较标准误与总体标准差可见，标准误明显较小。标准误为标准差的 $1/\sqrt{n}=1/\sqrt{\text{每份样本例数}}$ 倍。例如，某次实验获得标准误为 0.4571，与总体标准差 0.64 相差 $0.4571/0.64=1/\sqrt{1.96}\approx 1/\sqrt{2}$ 倍。

2. 直方图的绘制

为了更直观、动态地反映每次随机抽样实验后的样本均数的变化情况，对 1000 个样本均数，利用 FREQUENCY 函数产生频数表，并利用图表向导绘制直方图。具体操作如下：

①确定频数表的组段上限。

因为在正态分布中，（均数±2.58 标准差）将包括总体中 99%的个体。以 $(3.73-2.58\times0.64)=2.08$ 确定第一组段的上限，即在单元格 G2 处输入公式“=$L$2-2.58*$M$2”。假定分为 25 个组段（因为有 1000 个均数（不是个体值），分组可多一些），则组距应该等于[（均数+2.58×标准差）-（均数-2.58×标准差）]/25=2×2.58×标准差÷25，第二组段应该在第一组段的基础上加上组距。所以在单元格 G3 中应键入公式“=G2+(2*2.58*$M$2)/25”，将此公式向下拖放填充至单元格 G26。为了避免漏掉值，采用第 B 列的正态分布原数据获得最后一个组段上限值，即单元格 G27 中为“=MAX(B:B)”。

②用 FREQUENCY 函数产生频数表的频数。

选取单元格区域 H2:H27，在编辑栏输入公式“=FREQUENCY(D3:D1002,G2:G27)”，按 Ctrl+Shift+Enter 组合键（注意不要仅按 Enter 键）。在单元格区域 I2:I27 产生 H2:H27 所对应的频率，以便统一纵轴数据值在 0～1 之间。

③绘制直方图。

选择“插入→图表→柱形图”命令，单击第一个图表选项（簇状柱形图），在弹出的空白图表上单击鼠标右键，选择“选择数据”选项，在弹出的“选择数据源”对话框中，在“图例项”下单击“添加”按钮，在弹出的“编辑数据系列”中选择系列值范围为 I2：I27，单击“水平分类轴标签”下方的“编辑”按钮，选择轴标签区域为 G2:G27，单击“确定”按钮，可获得初始直方图。为了使图形好看，可以将鼠标移至任意一个直条上，双击选中全部直条，弹出“设置数据点格式”对话框，在系列选项中将分类间距调整成 0%，在填充中可选择图案填充，边框颜色可选择为实线。如果需要，可以单击直方图空白处，在弹出的图标工具栏中，选择设计、布局、格式等窗口，更改图表的布局、标签、坐标轴、背景等。修饰后的直方图如图 18-14 的右侧所示。

图 18-14 的左图是根据第 B 列原正态分布数据绘制的，反映了原数据的分布情况，制作方法与右图完全相同。



电脑实验

①打开光盘文件“18-1 正态分布总体抽样 s2.xlsx”文件，每按一次 F9 键，每份样本将重新被抽取，每份样本的均数与标准差、均数的均数与标准误、频数表频数、直方图将随之发生改变。

②改变正态分布参数总体均数与总体标准差，即改变单元格L2、M2中的数据，可观察不同正态分布总体的样本均数分布变化情况。

结果分析：从该实验可见，从正态总体中抽取样本，获得的样本均数的分布仍近似呈正态分布，即使样本含量很小（如 $n=2$ ）。样本均数的均数围绕总体均数上下波动，均数的标准差即标准误与总体标准差相差一个常数的倍数，即样本均数的标准误=总体标准差/$\sqrt{n}$，或 $\sigma_{\bar{X}}=\sigma/\sqrt{n}$ 。

如果增大每份样本的含量，如从 $n=2$，增加到 $n=5$ 、 $n=10$ 或 $n=30$，则抽样获得的标准误将更小，即获得的样本均数将更加集中，直方图更加“高瘦”（如图 18-15 所示）。

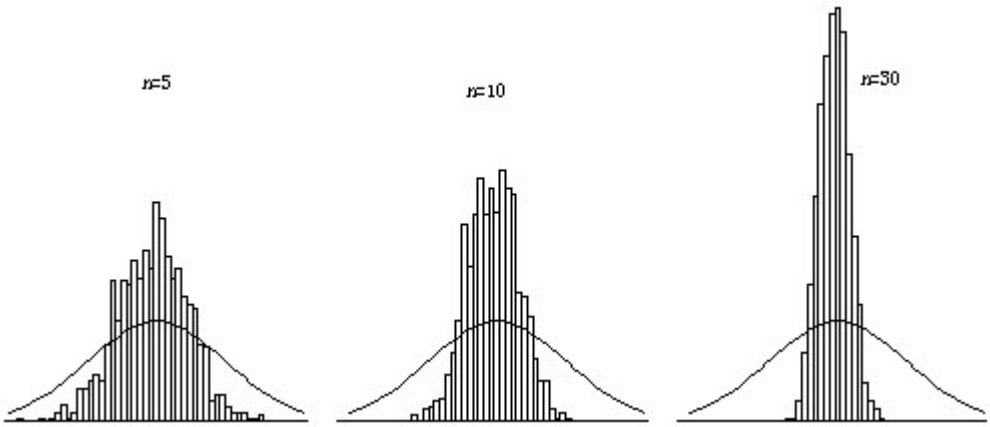


图 18-15　不同样本含量时的均数分布

在实际应用中，总体标准差 σ 往往未知，只能用样本标准差 S 代替 σ，从而获得 $\sigma_{\bar{X}}$ 的估计值：

$$S_{\bar{X}}=S/\sqrt{n} \qquad (18\text{-}1)$$

以示区别，将 $\sigma_{\bar{X}}$ 称为理论标准误， $S_{\bar{X}}$ 称为样本标准误。

温馨提示：在本书所附的光盘中，制作了 4 个 Excel 电脑实验“程序”（分别编号为 18-1、18-2、18-3、18-4），每个程序从正态总体中随机抽取 1000 份样本，每份样本分别含 $n=2$、5、10、30 个个体。

①打开 Excel 电脑实验“程序”，按 F9 键，观察每份样本被重新抽取后，样本的均数与标准差、均数的均数与标准误以及直方图的改变。

②比较不同样本含量时直方图、均数的均数及其标准误的改变情况。图 18-16 为 $n=30$ 的正态分布随机抽样电脑实验，试与图 18-14 进行比较。

③改变正态分布的总体参数（ μ,σ^2 ），再重复以上两点，可观察不同正态分布导致

的图形变化。一般情况下，改变正态分布参数，直方图图形变化不明显。

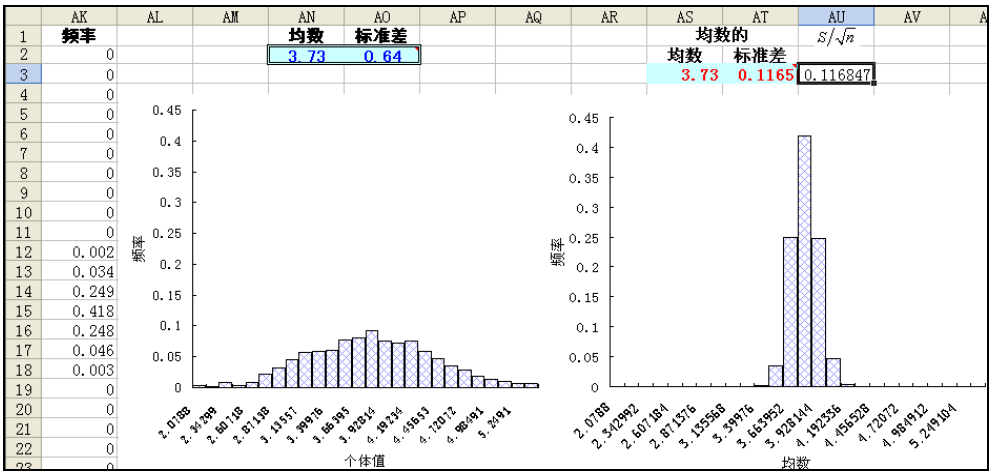


图 18-16　例数为 30 的电脑实验界面

18.3.2　从非正态总体中随机抽样

1. 均匀分布随机抽样数据的产生

均匀分布是一种典型的非正态分布。下面制作均匀分布抽样实验，进一步体会抽样误差的性质。

例 18-3：假定均匀分布的下限值为 1，上限值为 100，从这个总体中随机抽取 1000 份样本，每份含 n=2 个个体。求每份样本的均数和标准差，并对求得的 1000 个均数绘制直方图。

解：按如下步骤产生随机样本，并求均数与标准差。

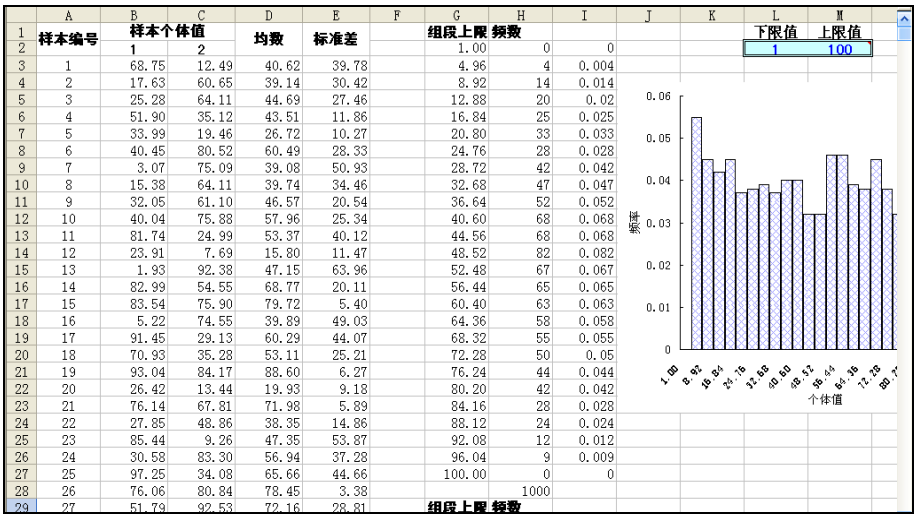
①按图 18-17（参见光盘文件“18-5 均匀分布总体抽样 s2.xlsx”），在单元格区域 A1:M1 等处分别键入并编辑文本或数字，其中单元格 L2、M2 中分别存放有均匀分布总体的下限值与上限值。

②在单元格 A3 中键入数字“1”，单击“开始→编辑→填充→系列”，在弹出的“序列”对话框中，“序列产生在”下选择“列”，并在“终止值”右侧框内输入“1000”，单击“确定”按钮，即可在单元格区域 A3:A1002 中产生 1～1000 的序列数。

③在单元格 B3 中输入公式“=$L$2+RAND()*($M$2–$L$2)”，将该公式向右拖放填充至单元格 C3，由此产生第一份样本。其中 L2、M2 分别为下限值、上限值。

④由第一份样本（B3:C3）计算样本均数与标准差。在单元格 D3 中键入公式

“=AVERAGE(B3:C3)”求均数；在单元格 E3 中键入公式“=STDEV(B3:C3)”求标准差。



样本编号	样本个体值		均数	标准差
	1	2		
1	68.75	12.49	40.62	39.78
2	17.63	60.65	39.14	30.42
3	25.28	64.11	44.69	27.46
4	51.90	35.12	43.51	11.86
5	33.99	19.46	26.72	10.27
6	40.45	80.52	60.49	28.33
7	3.07	75.09	39.08	50.93
8	15.38	64.11	39.74	34.46
9	32.05	61.10	46.57	20.54
10	40.04	75.88	57.96	25.34
11	81.74	24.99	53.37	40.12
12	23.91	7.69	15.80	11.47
13	1.93	92.38	47.15	63.96
14	82.99	54.55	68.77	20.11
15	83.54	75.90	79.72	5.40
16	5.22	74.55	39.89	49.03
17	91.45	29.13	60.29	44.07
18	70.93	35.28	53.11	25.21
19	93.04	84.17	88.60	6.27
20	26.42	13.44	19.93	9.18
21	76.14	67.81	71.98	5.89
22	27.85	48.86	38.35	14.86
23	85.44	9.26	47.35	53.87
24	30.58	83.30	56.94	37.28
25	97.25	34.08	65.66	44.66
26	76.06	80.84	78.45	3.38
27	51.79	92.53	72.16	28.81

组段上限	频数	
1.00	0	0
4.96	4	0.004
8.92	14	0.014
12.88	20	0.02
16.84	25	0.025
20.80	33	0.033
24.76	28	0.028
28.72	42	0.042
32.68	47	0.047
36.64	52	0.052
40.60	68	0.068
44.56	68	0.068
48.52	82	0.082
52.48	67	0.067
56.44	65	0.065
60.40	63	0.063
64.36	58	0.058
68.32	55	0.055
72.28	50	0.05
76.24	44	0.044
80.20	42	0.042
84.16	28	0.028
88.12	24	0.024
92.08	12	0.012
96.04	9	0.009
100.00	0	0
	1000	

下限值	上限值
1	100

图 18-17　均匀分布随机样本及每份样本的均数与标准差

⑤选取单元格区域 B3:E3，在被选取的单元格上单击鼠标右键，在弹出的快捷菜单中选择“复制”命令。选取单元格区域 B4:E1002（单击单元格 B4，用屏幕右侧的滚动条向下滚动直至看到单元格 E1002，按住 Shift 键，再单击单元格 E1002）。在常用工具栏中单击“粘贴”图标，便可产生 1000 份随机样本，并获得了每份样本（2 个个体值）的均数、标准差。

⑥在单元格区域 Q1:R2 中，按图 18-17 输入文字“均数的均数、标准差”。在单元格 Q3 中输入公式“=AVERAGE(D3:D1002)”，求 1000 份样本均数的均数；在单元格 R3 中输入公式“=STDEV(D3:D1002)”，求 1000 份样本均数的标准差。

只要光标在产生随机数的工作表中，按一次 F9 键，图 18-17 显示的随机样本及其每份样本的均数、标准差（即单元格区域 B3:E1002）、样本均数的均数与其标准差（单元格 Q3、R3）都会发生“随机”的改变。

2. 均匀分布样本均数直方图的绘制

为了更直观地反映每次随机抽样获得的样本均数的分布，对 1000 个样本均数，利用 FREQUENCY 函数产生频数表，利用图表向导绘制直方图。具体操作如下：

①确定频数表的组段上限。

均匀分布有下限值和上限值，可分别当作其最小值与最大值。假定分为 25 个组段，则组距应该等于“（上限值-下取值）/25”。因此，令第一组段的上限值等于下限值，即 Excel 公式为“=$L$2”；以后每一组段应在上一组段的基础上加组距，如在单元格 G3 中应键入第二组段公式“=G2+($M$2-$L$2)/25”，将此公式向下拖放填充至单元格 G27。此时单元格

G27 中的 Excel 公式应该是“=G26+($M$2–$L$2)/25”。

②用 FREQUENCY 函数产生频数表的频数。

选取单元格区域 H2:H27，在编辑栏输入公式“=FREQUENCY(D3:D1002,G2:G27)”，按 Ctrl + Shift + Enter 组合键即可（注意不要仅按 Enter 键）。在 I2:I27 处，计算单元格区域 H2:H27 所对应的频率，以便统一纵坐标的尺度在 0～1 范围。

- 绘制直方图。

单击插入图表中柱形图，选择二维的簇状柱形图，在弹出的空白图表上，单击鼠标右键选择“选择数据”命令，在弹出的“选择数据源”对话框中，在“图例项”下单击“添加”按钮，在弹出的编辑数据系列中选择系列值范围为 I2:I27，单击“水平分类轴标签”下方的“编辑”按钮，选择轴标签区域为 G2:G27，单击“确定”按钮，可获得初始直方图。在“图标工具”窗口下方的标签中，选择坐标轴标题，分别添加横坐标标题“均数”和纵坐标标题“频率”，再经过适当修饰可完成直方图绘制。

3. 均匀分布原个体值直方图的绘制

均匀分布总体的样本均数也服从正态分布。为了比较原个体值随机数分布与样本均数分布的差别，按下列方法绘制原个体值随机数直方图。

①将单元格区域 G1:I28 的内容复制、粘贴到单元格区域 G29:I56。

②选中单元格区域 H30:H55，在编辑栏输入公式“=FREQUENCY(B3:B1002,G30:G55)”，按 Ctrl + Shift + Enter 组合键，获得第 B 列的频数。

③原个体值随机数分布（左）直方图的制作与前面相同，制作好的图形如图 18-18 所示。

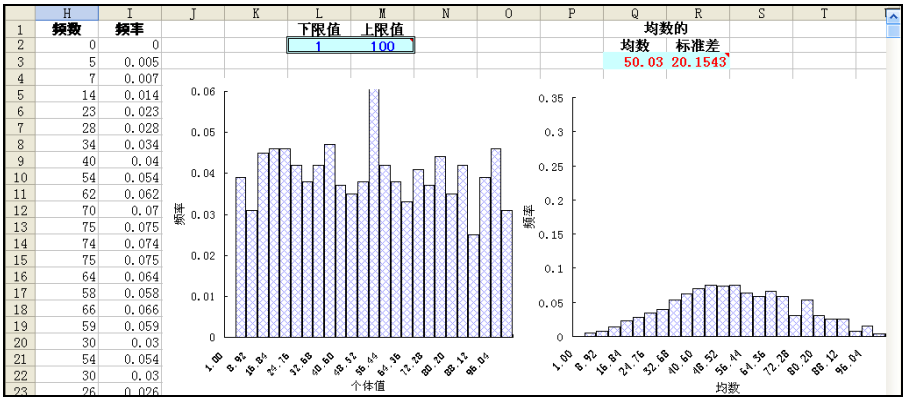


图 18-18 原个体值分布（左）与样本均数分布（右）的直方图

电脑实验

①打开光盘文件“18-5 均匀分布总体抽样 s2.xlsx”文件，每按一次 F9 键，每份样本

将重新被抽取，每份样本的均数与标准差、均数的均数与标准误、频数表频数、直方图将随之发生改变。

②改变均匀分布的下限值与上限值，即改变单元格 L2、M2 中的数据，可观察不同均匀分布总体的样本均数分布变化情况。

结果分析：从该实验可见，从非正态总体中抽取样本，在样本含量足够大的情况下，获得的样本均数分布仍近似呈正态分布。当样本含量很小（如 n=2）时，样本均数的分布略有偏离正态分布的情况。随着样本含量的增大，样本均数的分布会逐步趋近于正态分布。样本含量足够大一般是指样本含量大于等于 30，但 30 也不是恒定的标准。如果非正态分布（如均匀分布）近似于正态分布，样本含量可少些；如果非正态分布（如双峰分布）与正态分布相距较远，则样本含量应更大些。

从这一电脑实验可见：

①即使从非正态总体中抽取样本，所得样本均数的分布仍近似呈正态分布。

②随着样本含量的增大，样本均数的变异范围逐渐变窄。

温馨提示：在本书所附的光盘中，制作了 4 个 Excel 电脑实验“程序”（分别编号为 18-5、18-6、18-7、18-8），每个程序从正态总体中随机抽取 1000 份样本，每份样本分别含 n=2、5、10、30 个个体。

①打开Excel电脑实验“程序”，按F9键，观察每份样本被重新抽取后，样本的均数与标准差、均数的均数与标准误以及直方图的改变。

②比较不同样本含量时直方图、均数的均数及其标准误的改变情况。图18-19为n=30的正态分布随机抽样电脑实验，试与图18-18进行比较。

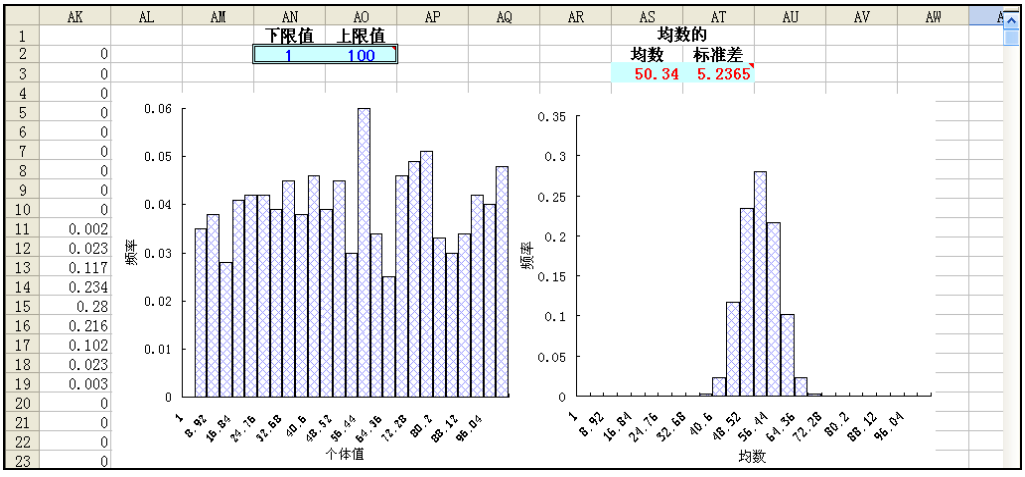


图 18-19　原个体值随机数分布（左）与样本均数分布（右）的直方图

以上实验的现象有人总结称之为中心极限定理（central limit theorem），即在样本含量足够大的情况下，从任何总体中抽取的随机样本之均数的抽样分布将近似服从正态分布。样本含量越大，抽取的随机样本均数的抽样分布越逼近正态分布。原样本是正态分布，无论样本含量多大，样本均数的抽样分布均服从正态分布；原样本为非正态分布，样本含量足够大时样本均数的抽样分布也服从正态分布，一般认为样本含量30及以上为足够大。但总体如果是极度非正态（如双峰分布、极度偏态分布），那么在适度样本含量情况下样本均数的抽样分布也有可能不服从正态分布。

第 19 章

假设检验的概率分布与统计用表

在第 8、9 两章中提到了 z 检验、t 检验、F 检验（方差分析）、χ^2 检验。它们对应的概率分布分别是 z 分布、t 分布、F 分布、χ^2 分布，它们都与正态分布有密切关系，都属于连续分布。

19.1 z 分布

当总体均数 μ=0，总体标准差 σ=1 时，一般正态分布就变成了标准正态分布，即 z 分布，简记为 $N(0,1)$，其图形见第 17 章图 17-1 的正态曲线 $f(X)$。

将 μ=0、σ=1 代入第 17 章公式（17-1），可得到

$$f(X)=\frac{1}{\sqrt{2\pi}}e^{-\frac{X^2}{2}}$$

为了和其他一般正态分布相区分，记横坐标值 X 为 z，纵坐标值 $f(X)$ 为 $\varphi(z)$，从而标准正态分布的概率密度函数为：

$$\varphi(z)=\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}},\quad -\infty<z<+\infty \tag{19-1}$$

该公式对应的 Excel 函数格式为“=NORMDIST(z,0,1,0)，即=NORMDIST(分位值 z, 均数 0,标准差 1,是否累积的指示 0)”，Excel 2010 版本增加了“=NORM.DIST(z,0,1,0)”函数。对于一般正态分布资料，均可采用

$$z=\frac{X-\mu}{\sigma} \tag{19-2}$$

变换成标准正态分布。

标准正态分布的累积分布函数为：

$$\Phi(Z=z)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z}e^{-\frac{Z^2}{2}}dZ,\quad -\infty<Z<+\infty \tag{19-3}$$

该公式对应的 Excel 函数格式为“=NORMSDIST(z)”。Excel 2010 版本增加了“=NORM.S.DIST(z,1)”函数，后面的 1 为累积分布函数，0 为概率密度函数。

19.1.1 标准正态分布曲线下面积

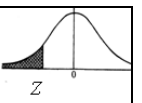
给定一个 z 值或一个 z 值范围，便可通过 Excel 函数“=NORMSDIST(z)”获得正态分布曲线下面积。一般统计教科书给出的“标准正态分布 z 值表”（如图 19-1 所示）就是根据公式（19-3）制定的。例如 z=–2.88 时，Excel 函数“=NORMSDIST(–2.88)”的结果为“0.0020”；z=–2.09 时，Excel 函数“=NORMSDIST(–2.09)”的结果为“0.0183”。

由此可见，图 19-1 的 z 值表是计算 z 分布曲线下从 –∞ 到 z，即 z 值以左的曲线下面积；如果要计算 z 值以右（即从 z 到+∞）的曲线下面积，则 Excel 函数为“=1–NORMSDIST(z)”。

曲线下面积具有一定的规律性，例如正态分布曲线下均数左右 1 倍标准差所夹面积为

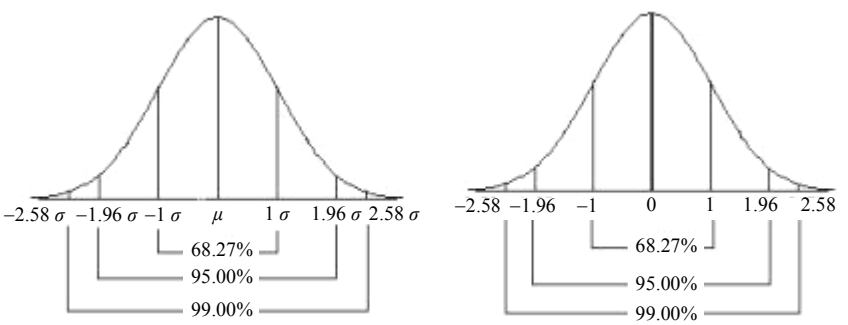
68.27%（Excel 函数“=NORMSDIST(1)–NORMSDIST(–1)”）；均数左右 1.96 倍标准差所夹面积为 95.00%（Excel 函数“=NORMSDIST(1.96)–NORMSDIST(–1.96)”）；均数左右 2.58 倍标准差所夹面积为 99.00%（Excel 函数“=NORMSDIST(2.58)–NORMSDIST(–2.58)”），如图 19-2 所示。

附表 1　　z 值表（标准正态分布曲线下的面积）$\Phi(Z=z)$ 值



Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
−3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
−2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
−2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
−2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
−2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
−2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
−2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
−2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
−2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
−2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
−2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183

图 19-1　z 值表



（a）正态分布的面积规律　　（b）标准正态分布的面积规律

图 19-2　正态分布与标准正态分布的面积规律

在许多医学研究中，常对一组正常人的某项指标进行测定，以确定该指标的参考值范围（reference range）。当某项指标测量值的直方图接近正态分布时，可利用上述结果估计总体的 95%或 99%范围。μ 和 σ 通常未知，只能采用样本均数 $\overline{X}$ 和样本标准差 S 来代替，以（$\overline{X}-1.96S, \overline{X}+1.96S$）和（$\overline{X}-2.58S, \overline{X}+2.58S$）作为该项指标的 95%和 99%参考范围。

19.1.2 标准正态分布分位数

标准正态分位数即累积分布函数的反函数。如果已知标准正态分布曲线下左侧面积 $\Phi(z)=\alpha$，便可反推出标准正态分位数 z。横坐标相应的 z 值就是分位数（quantile）或临界值，常用 z_α 表示。如 $\Phi(z_\alpha)=\alpha=95\%$，表明横坐标值为 z_α 的曲线下左侧面积为 95%。$z=\Phi^{-1}(95\%)=\Phi^{-1}$（曲线下左侧面积）$=\Phi^{-1}(\alpha)$。

z 值的 Excel 函数格式为：NORMSINV(概率)。如果为一般正态分布，其 Excel 函数格式为：NORMINV(概率，均数，标准差)。

例 19-1：求单侧检验和双侧检验 $\alpha=0.05$ 的 z 界值。

解：对于单侧检验，有 Excel 公式“=NORMSINV(0.05)”，即单侧 $z_{0.05}=-1.6449$。对于双侧检验，表示正态分布曲线中间占 95%，两端分别占 2.5%，Excel 公式应为“=NORMSINV(0.025)”，即双侧 $z_{0.05/2}=-1.96$。

19.1.3 标准正态分布的其他应用

（1）二项分布的总体率置信区间

当样本含量 n 较大，样本率 p 不太小或不太大，以至 np 和 $n(1-p)$ 均大于 5 时，可采用近似正态分布 $N\left(np,np(1-p)\right)$ 求总体率 π 的 $100(1-\alpha)\%$ 置信区间如下：

$$\left(p-z_{\alpha/2}S_p,p+z_{\alpha/2}S_p\right) \qquad (19\text{-}4)$$

设样本率 $p=25\%$ 和 $S_p=0.0306$，并分别放在单元格 A1、B1 中。单击单元格 D1，然后在编辑栏输入“=$A$1-NORMSINV(0.975)*$B$1”；将单元格 D1 中公式拖放填充至 D2，并修改为“=$A$1+NORMSINV(0.975)*$B$1”。可获得 $\alpha=0.05$，95%置信区间为（5.75%,11.87%）。由 NORMSINV(0.975)获得标准正态分位数 $z_{0.05/2}$，如果求 99%置信区间应改为 NORMSINV(0.995)。

（2）二项分布的样本率与总体率比较

样本率与总体率的比较涉及第 17 章的二项分布知识，但为了连贯性，暂且将应用实例放在此处。

例 19-2：据有关文献报道，新生儿染色体异常率为 1%，某研究者随机抽查当地 500 名新生儿，结果 1 名新生儿染色体异常，问按二项分布原理该地新生儿染色体异常是否低于一般水平？

解：已知 $n=500$，$X=1$，$\pi=1\%=0.01$；问该地新生儿染色体异常是否低于一般水平，相当于问在二项分布 $B(500,0.01)$ 中小于等于 1 的概率有多大，即求 $P(X\leqslant 1)$。

单击任一单元格，如 A1，在编辑栏输入二项分布累计分布函数公式“=BINOMDIST(1, 500, 0.01, 1)”，单击“确定”得 $P(X\leqslant 1)=0.0398$。如果按检验水准 $\alpha=0.05$ 判断，则获得的差异具有统计学意义。这就是说有理由认为该地新生儿染色体异常低于一般水平。

（3）二项分布的两样本率比较

例 19-3：某医生欲了解螺赛克治疗消化道溃疡，以雷米替叮作对照组，试验结果如表 19-1 所示。问两种药物治疗 4 周的疗效有无差别？

表 19-1　两种药物治疗消化道溃疡效果

处　理	治　愈	未 治 愈	合　计	治愈率(%)
螺赛克	44(X_1)	11(n_1-X_1)	55(n_1)	80.00(p_1)
雷米替叮	35(X_2)	13(n_2-X_2)	48(n_2)	72.92(p_2)
合计	79	24	103	

解：当计算两样本率的样本含量较大，样本率与 0.5 相差不远，以至 n_1p_1、$n_1(1-p_1)$、n_2p_2、$n_2(1-p_2)$ 均大于 5（如本例）时，比较两样本率可采用如下 z 统计量：

$$z=\frac{p_1-p_2}{S_{p_1-p_2}},\quad S_{p_1-p_2}=\sqrt{\frac{X_1+X_2}{n_1+n_2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} \tag{19-5}$$

$$S_{p_1-p_2}=\sqrt{\frac{44+35}{55+48}\left(\frac{1}{55}+\frac{1}{48}\right)}=0.1730,\quad z=\frac{0.8000-0.7292}{0.1730}=0.4095$$

对于单侧检验，由 z 统计量获得检验 P 值的 Excel 公式为：

“=1-NORMSDIST(0.4095)”

得 P=0.3411。

对于双侧检验，由 z 统计量获得检验 P 值的 Excel 公式为：

“=2*(1–NORMSDIST(0.4095))”

得 P=0.6822。在 $\alpha=0.05$ 检验水准下，两种药物治疗 4 周后的治愈率差异无统计学意义。

（4）Poisson 分布的总体均数置信区间

当样本计数 X 较大（如 $X>20$）时，可利用 Poisson 分布的正态近似性，计算其总体均数 100($1-\alpha$)%置信区间如下：

$$\left(X-z_{\alpha/2}\sqrt{X},\ X+z_{\alpha/2}\sqrt{X}\right) \tag{19-6}$$

式中，X 为样本计数；$z_{\alpha/2}$ 为 α 检验水准双侧标准正态分布分位数，如 $z_{0.05/2}=1.96$。

（5）Poisson 分布的样本均数与总体均数比较

Poisson 分布的样本均数与总体均数的比较，涉及第 17 章的 Poisson 分布知识，为了连贯性，暂且将应用实例放在此处。

例 19-：据以往大量观察得知，某溶液中平均每毫升有 4 个细菌。某研究者采取已放在 5℃冰箱中 3 天的该溶液 1 毫升测得 9 个细菌。问放在 5℃冰箱中 3 天的溶液中细菌数是否有增长？

解：1 毫升测得 9 个细菌，是单位容积内计数资料，服从 Poisson 分布。问该溶液中细菌数是否有增长，即求 $P(X\geqslant 9)$是否小于检验水准α。因为 $P(X\geqslant 9)=1-P(X\leqslant 9)+P(X=9)$，且参数 $\lambda=4$，所以其对应的 Excel 公式为：“=1–POISSON(9,4,1)+POISSON(9,4,0)”，得 $P(X\geqslant 9)=0.0214$。在$\alpha=0.05$ 检验水准下，可以认为放在 5℃冰箱 3 天的溶液中细菌数有增长。

（6）Poisson 分布的两样本均数比较

例 19-5：某环境监测员对两种水样各抽验 1ml，分别培养出大肠杆菌 13 个和 20 个，试比较这两种水样中平均每毫升所含大肠杆菌数有无差别？

解：本例 $X_1=13$，$X_2=20$，$(X_1+X_2)>20$，可采用正态分布近似方法，即按公式：

$$z=\frac{X_1-X_2}{\sqrt{X_1+X_2}} \tag{19-7}$$

本例为：

$$z=\frac{13-20}{\sqrt{13+20}}=-1.2185$$

由 z 统计量获得单侧检验 P 值的 Excel 公式为：“=NORMSDIST(–1.2185)”，P=0.1116；由 Excel 公式“=2*NORMSDIST(–1.2185)”，获得双侧检验 P=0.2232。表明在$\alpha=0.05$ 检验水准下，尚不能认为两种水样平均每毫升所含大肠杆菌数有差别。

例 19-6：用核素方法独立地测量两份标本的放射性，标本制备完全相同，但测量时间长短不一。对于标本 1，10 分钟测得质点数为 2000；对于标本 2，18 分钟测得质点数为 4500。问若在相同时间长度内测量，两份标本发放质点是否有差异？

解：对应不同容量、体积、时间等单位的数据，进行 Poisson 分布的两种样本均数的比较时，计算公式与式（19-7）略有不同。本例两份标本测量时间有差异，令 $n_1=10$，$X_1=2000$，$n_2=18$，$X_2=4500$，有$\left(\frac{X_1}{n_1}+\frac{X_2}{n_2}\right)>20$，可采用正态分布近似方法，即按公式计算：

$$z=\frac{\frac{X_1}{n_1}-\frac{X_2}{n_2}}{\sqrt{\frac{X_1}{n_1^2}+\frac{X_2}{n_2^2}}} \tag{19-8}$$

本例为：

$$z=\frac{\frac{2000}{10}-\frac{4500}{18}}{\sqrt{\frac{2000}{10^2}+\frac{4500}{18^2}}}=-8.5890$$

由 z 统计量获得双侧检验 P 值的 Excel 公式为“=2*NORMSDIST(–8.5890)”，得 P=8.7730E–18。表明在 $\alpha=0.05$ 检验水准下，可认为两份标本发放质点有差异。

19.1.4 z 分布电脑实验

（1）数据准备

按如图 19-3 所示（可见光盘文件“19-1 z 分布.xlsx”）在单元格 A3 中键入“-4”，单元格 A4 中键入“–3.99”，选取 A3:A4 单元格区域，将此向下拖放至 A803，产生 z 分布横坐标 z 值；在单元格 B3 中键入“=NORMDIST(A3,0,1,0)”，将此公式向下拖放至 B803，产生变量 z 值所对应的 z 分布纵坐标 $\varphi(z)$ 值；为了反映 z 值以左曲线下面积，在单元格 C3 中键入“=IF($F$5<=A3,0,B3)”，将此公式向下拖放至 C803，使大于等于 z 值的曲线下纵坐标变为 0；为了反映 z 值以右曲线下面积，在单元格 D3 中键入“=IF($F$5>A3,0,B3)”，将此公式向下拖放至 D803，使小于 z 值的曲线下纵坐标变为 0。

（2）滚动条

按如图 19-3 所输入单元格区域 F3:L4 中的文字，左滚动条（由“开发工具→插入→ActiveX 控件→滚动条”产生）的属性规定 max 为 800，min 为 0，linkedcell 为 F6，在 F5 单元格中键入公式“=(F6–400)/100”，使滚动条在–4 至 4 之间变化，步幅为 0.01，即和第 A 列 z 变量取值相同。为了获得单元格 F5 中 z 值对应的 P 值，在单元格 G5、H5 中分别键入公式“=(1–NORMSDIST(ABS($F5)))”、“=(1–NORMSDIST(ABS($F5)))*2”，获得单侧、双侧 P 值。该结果与图 19-1（z 值表）完全对应。

右滚动条（由“开发工具→插入→ActiveX 控件→滚动条”产生）的属性规定 max 为 500，min 为 0，linkedcell 为 J6，在 J5 单元格中键入公式“=J6*0.001”，使滚动条在 0 至 0.5 之间变化，步幅为 0.001。为了获得单元格 J5 中 P 值对应的 z 值，在单元格 L5 中键入公式“=NORMSINV(J5)”，获得单侧 α 对应的 z 临界值。单元格 K5 是 z 临界值对应的双侧 α，即单侧 α 的 2 倍。

（3）绘制 z 分布图

由 Excel 面积图第一个子图绘制 z 分布图（如图 19-3 左侧所示）。首先以单元格区域 A3:A803 为横坐标值，B3:B803 为纵坐标值，绘制整个 z 分布图；在同一图中，然后以单元格区域 A3:A803 为横坐标值，C3:C803 为纵坐标值，绘制单元格 F5 对应 z 值以左的 z 分布图。这样滑动左滚动条，单元格 F5 中 z 值发生变化，z 分布图也相应发生变化。

图 19-3 右侧 z 分布图绘制方法类似，只是由单元格区域 D3:D803 数据替代了 C3:C803。

该图反映了标准正态分布曲线下 z 值以右曲线下面积随单元格 F5（由滚动条控制）的变化情况。

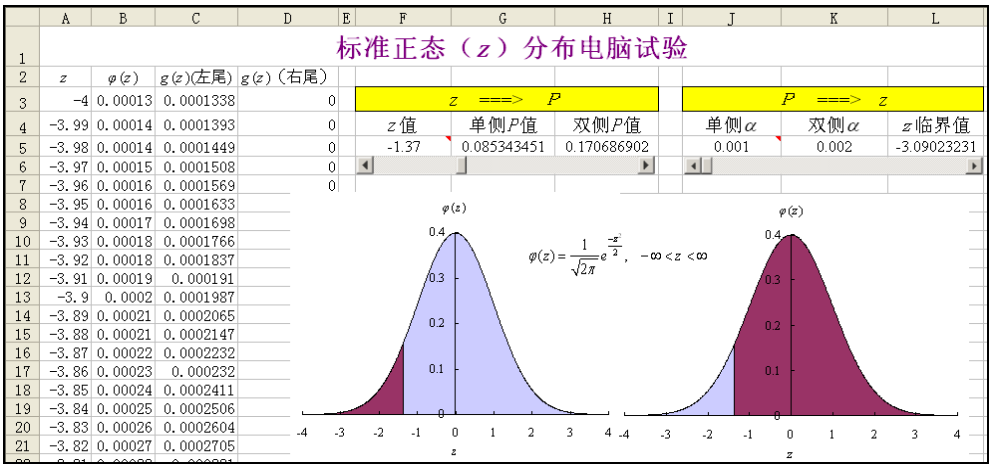


图 19-3 标准正态分布的电脑实验

电脑实验

打开光盘中“19-1 z 分布.xlsx”文件，在位于单元格区域 F6:H6 处，左右滑动滚动条，可以发现：

①当滚动条向左滑动时，单元格 F5 中的 z 值逐渐变小；向右滑动时，单元格 F5 中的 z 值逐渐变大，z 在–4 至+4 之间变化，步幅为 0.01。

②z 值对应的单侧（单元格 G5）、双侧（单元格 H5）P 值也相应发生改变。

③当滚动条向左滑动时，左侧 z 分布图深色阴影向左移动；当滚动条向右滑动时，右侧 z 分布图紫红色阴影向右移动，反映了 z 分布曲线下 z 值以左的曲线下面积 $P(Z\leqslant z)$随 z 值的变化而变化；右侧 z 分布图深色阴影同时发生相应变化，反映了 z 分布曲线下 z 值以右的曲线下面积 $P(Z>z)$随 z 值的变化而变化。

19.2 t 分布

第 18 章的电脑实验已经表明：从正态总体 $N(\mu,\sigma^2)$ 中随机抽取样本，获得的样本均数 $\overline{X}$ 服从正态分布 $N(\mu,\sigma^2/n)$。样本均数 $\overline{X}$ 的标准正态离差为：

$$z=\frac{\overline{X}-\mu}{\sigma_{\overline{X}}}=\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \qquad (19\text{-}9)$$

温馨提示：①为了与“由 z 值获得 P 值”形成对照，在该文件的单元格区域 J6:L6 处左右滑动右侧滚动条，可以“由 P 值（即 α 值）获得 z 临界值”。

②在“19-1 z 分布.xlsx”文件的下方，有一个图可同时修改、观看 z 分布曲线下 z 值以左与 z 值以右曲线下面积（如图 19-4 所示），其相应数据分别位于第 A、B、N、O 列。

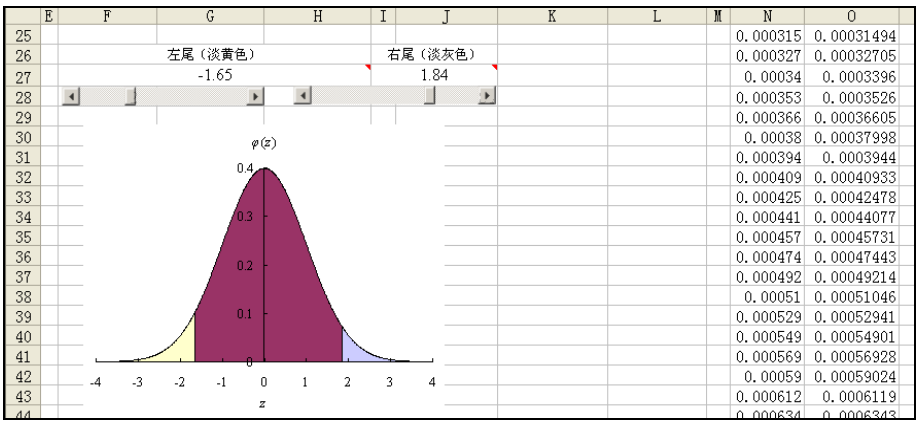


图 19-4　标准正态分布的双侧尾部面积的改变

服从标准正态分布，即 z 分布。σ在通常情况下未知，一般采用样本标准差 S 代替公式（19-9）中的总体标准差σ，从而有

$$t=\frac{\overline{X}-\mu}{S_{\overline{X}}}=\frac{\overline{X}-\mu}{S/\sqrt{n}},\quad \nu=n-1 \tag{19-10}$$

显然，$(\overline{X}-\mu)/S_{\overline{X}}$ 不同于 $(\overline{X}-\mu)/\sigma_{\overline{x}}$，不再服从 z 分布。例如，从某已知的正态总体中随机抽取 1000 份样本含量为 n 的样本，理论标准误是常量 $\sigma_{\overline{X}}=\sigma/\sqrt{n}$，而样本标准误却有 1000 个，即 $S_{\overline{X}}=S_1/\sqrt{n},\ S_2/\sqrt{n},\ \ldots\ldots,\ S_{1000}/\sqrt{n}$，它们不尽相同，即 $S_{\overline{X}}$ 有变异。Gosett WS 于 1908 年用笔名 Student 研究了它的分布规律，并称之 t 分布（t distribution）。公式（19-10）中的 ν（读作“nu”，为希腊字母）是 t 分布的自由度，不同的自由度对应于不同的 t 分布。ν 为 t 分布的唯一参数，所以记 t 分布为 $t(\nu)$。

19.2.1　t 分布的图形及其分布特征

t 分布的概率密度函数为：

$$f(t)=\frac{\Gamma[(\nu+1)/2]}{\sqrt{\pi\nu}\,\Gamma(\nu/2)}\left(1+\frac{t^2}{\nu}\right)^{-(\nu+1)/2},\quad -\infty<t<+\infty \tag{19-11}$$

式中，Γ(•) 为伽玛函数（Excel 中只有伽玛函数的自然对数函数，经变换可求得）；π 为圆周率（Excel 函数为 PI()）；ν 为自由度；t 为随机变量。以 t 为横轴，$f(t)$ 为纵轴，可绘制 t 分布图形。

例 19-7：请在同一个图形中绘制 $t(1)$、$t(5)$ 两个 t 分布和一个 z 分布图形。分析比较 t 分布图形的特点。

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	t	t分布1	t分布2	正态分布			t分布1		t分布2			正态分布	
2	-4.0	0.018724	0.005124	0.000134			自由度1(v1)		自由度2(v2)			均数	标准差
3	-3.9	0.019637	0.005748	0.000199			1		5			0	1
4	-3.8	0.020616	0.006459	0.000292									
5	-3.7	0.021668	0.007268	0.000425									
6	-3.6	0.022802	0.008191	0.000612									
7	-3.5	0.024023	0.009244	0.000873									
8	-3.4	0.025343	0.010449	0.001232									
9	-3.3	0.026771	0.011827	0.001723									
10	-3.2	0.028319	0.013406	0.002384									
11	-3.1	0.030001	0.015216	0.003267									
12	-3.0	0.031831	0.017293	0.004432									
13	-2.9	0.033827	0.019677	0.005953									
14	-2.8	0.036008	0.022416	0.007915									
15	-2.7	0.038397	0.025562	0.010421									
16	-2.6	0.041019	0.029176	0.013583									
17	-2.5	0.043905	0.033326	0.017528									
18	-2.4	0.047087	0.038090	0.022395									
19	-2.3	0.050606	0.043551	0.028327									
20	-2.2	0.054505	0.049803	0.035475									
21	-2.1	0.058837	0.056948	0.043984									
22	-2.0	0.063662	0.065090	0.053991									
23	-1.9	0.069048	0.074342	0.065616									
24	-1.8	0.075073	0.084813	0.078950									

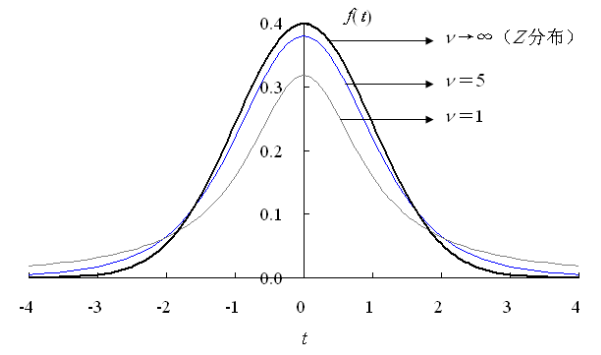


图 19-5 t 分布与标准正态分布曲线

解：图 19-5 对应的光盘文件为“19-2 t 分布 1.xlsx”。制作步骤如下：

① 在单元格 A2 中输入−4，用鼠标选中 A2，单击功能区中的“开始→编辑→填充→系列”命令，在“序列”对话框中选择“序列产生在列”、“等差序列”，“步长值”、“终止值”分别填入 0.1、4，产生第 A 列随机变量 t 值的数据。按如图 19-5 所示在单元格区域 G1:M3 中键入相应文字或数据。

② 为了产生随 t 值变化的 t 分布 1 的纵坐标值，按公式（19-11）在单元格 B2 中键入 Excel 公式：

“=EXP(GAMMALN(($G$3+1)/2))/SQRT($G$3*PI())/EXP(GAMMALN($G$3/2))*
(1+$A2^2/$G$3)^(($G$3+1)/2*(-1))”

为了产生随 t 值变化的 t 分布 2 的纵坐标值，按公式（19-11）在单元格 C2 中键入 Excel 公式：

“=EXP(GAMMALN(($I$3+1)/2))/SQRT($I$3*PI())/EXP(GAMMALN($I$3/2))*
(1+$A2^2/$I$3)^(($I$3+1)/2*(-1))”

为了产生随“t”值变化时标准正态分布的纵坐标值，在单元格 D2 中输入 Excel 公式：

“=NORMDIST(A2,$L$3,$M$3,0)”

（也可采用 NORMSDIST 函数。这里采用 NORMDIST 函数，目的是为了能改变均数标准差，并能绘制更多类型的正态分布。）

③选取并复制单元格区域 B2:D2，单击单元格 B3，移动右侧垂直滚动条至看到单元格 D82 为止，按住 Shift 键，用鼠标单击单元格 D82，在选中单元格区域 B3:D82 的情况下，单击工具栏中的“粘贴”，即可将单元格区域 B2:D2 中的三个公式复制填充到单元格区域 B3:D82 中。

④获得数据后，可利用这些数据及图表向导中的 *XY* 散点图绘制图形。

单击“插入→图表→散点图”图标的下拉窗口，选取子图表类型中的“带平滑线的散点图”。在弹出的图表中，在空白处单击鼠标右键，在弹出的选项中单击“选择数据”，弹出“选择数据源”对话框，选择图表数据区域$A$1:$D$82，单击“确定”按钮后即得到 *t* 分布图。为了使获得的图形美观好看，可以通过“图标工具”清除图域的背景色、网格线，改变横轴、纵轴刻度，添加标题横轴（*t*）、纵轴标题（*f*(*t*)），改变曲线颜色和样式，调整图例位置等。修改后的曲线图如图 19-5 所示。为了便于辨认，将存放各曲线名称、均数、标准差及 *t* 分布自由度等的文字、数据颜色改变成与曲线一致。

电脑实验

反复修改 *t* 分布的自由度，观察比较不同自由度的 *t* 分布曲线，体会 *t* 分布的如下性质：

①为单峰曲线分布，曲线在 $t=0$ 处最高；

②以 $t=0$ 为中心左右对称；

③*t* 分布曲线形状随自由度的不同而发生变化。自由度越小，$t=0$ 处的曲线纵高越矮，两尾部翘得越高；

④*t* 分布的极限为标准正态分布。随着自由度 ν 的增大，*t* 分布曲线逐渐逼近于标准正态分布，当 $\nu=\infty$ 时，*t* 分布曲线将完全与标准正态曲线重叠。

19.2.2 *t* 分布曲线下面积与检验界值

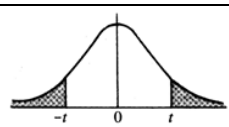
t 分布类似于标准正态分布，曲线下面积为 1。可用 Excel 的 TDIST 函数计算 *t* 分布曲线下面积。TDIST 函数的语法格式为“=TDIST(*t* 值，自由度，1 或 2)”。选“1”时返回 *t* 分布单侧右尾部的面积；选“2”时返回 *t* 分布双侧尾部的面积。同样，也可根据面积求 *t* 界值，其函数的语法格式为“=TINV(右侧尾部面积，自由度)”。

例如 *t* 值为 1.96，自由度为 20 时，采用 Excel 公式“=TDIST(1.96,20,1)”得单侧尾部面积为 0.0320；采用 Excel 公式“=TDIST(1.96,20,2)”得双侧尾部面积为 0.0641。标准正态

分布 z 值为 1.96 所对应的单侧尾部面积为 0.025，双侧尾部面积为 0.05。与之相比，对应于相同大小的变量值，t 分布所对应的 P 值比正态分布要大。

如果已知自由度为 20，规定检验水准 $\alpha=0.05$（即右侧尾部面积），采用 Excel 公式“=TINV(2*0.05, 20)”得单侧检验 t 界值为 1.7247；采用 Excel 公式“=TINV(0.05, 20)”得双侧检验 t 界值为 2.0860（见图 19-6）。检验水准 $\alpha=0.05$ 时，正态分布的单侧界值为 1.64，双侧界值为 1.96。与之相比，对应于同样大小的尾部面积 α，t 界值比正态分布界值要大。

附表 2 t 界值表



自由度		概率，P									
ν	单侧：	0.25	0.20	0.10	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
	双侧：	0.50	0.40	0.20	0.10	0.05	0.02	0.01	0.005	0.002	0.001
1		1.000	1.376	3.078	6.314	12.706	31.821	63.657	127.321	318.309	636.619
2		0.816	1.061	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.599
3		0.765	0.978	1.638	2.353	3.182	4.541	5.841	7.453	10.215	12.924
4		0.741	0.941	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5		0.727	0.920	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6		0.718	0.906	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7		0.711	0.896	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8		0.706	0.889	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9		0.703	0.883	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10		0.700	0.879	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11		0.697	0.876	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12		0.695	0.873	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13		0.694	0.870	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14		0.692	0.868	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15		0.691	0.866	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16		0.690	0.865	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17		0.689	0.863	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18		0.688	0.862	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19		0.688	0.861	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20		0.687	0.860	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850

图 19-6 t 界值表

19.2.3 t 分布电脑实验

t 分布电脑实验存放在光盘文件“19-3 t 分布 2.xlsx”中，具体制作步骤如下：

（1）数据准备

在单元格 A3 中键入“–4”，在单元格 A4 中键入“–3.99”，选取 A3:A4 单元格区域，将此向下拖放至 A803，产生 t 分布横坐标 t 值；在单元格 B3 中键入“=EXP(GAMMALN(($G$2+1)/2))/SQRT($G$2*PI())/EXP(GAMMALN($G$2/2))*(1+A3^2/$G$2)^(($G$2+1)/2*(–1))”，将此公式向下拖放至 B803，产生变量 t 值所对应的 t 分布纵坐标 $f(t)$ 值；为了反映 t 值以左曲线下面积，在单元格 C3 中键入“=IF($F$6<=A3,0,B3)”，将此公式向下拖放至 C803，使大于等于 t 值的曲线下纵坐标变为 0；为了反映 t 值以右曲线下面积，在单元格 D3 中键入“=IF($F$6>A3,0,B3)”，将此公式向下拖放至 D803，使小于 t 值的曲线下纵坐标变为 0。

（2）滚动条

t 分布滚动条的制作与 *z* 分布类似，不同的是 *t* 分布多了一个自由度滚动条。自由度滚动条的属性规定 max 为 200，min 为 1，linkedcell 为 G2，在单元格 G2 中的自由度可在 1 至 200 之间变化，步幅为 1。下方左滚动条的属性规定 max 为 800，min 为 0，linkedcell 为 F7，在 F6 单元格中键入公式“=(F7–400)/100”，使滚动条在-4 至 4 之间变化，步幅为 0.01，即和第 A 列 *t* 变量取值相同。为了获得单元格 F6 中 *t* 值对应的 *P* 值，在单元格 G6、H6 中分别键入公式“=TDIST(ABS($F$6),$G$2,1)”、“=TDIST(ABS($F$6),$G$2,2)”，获得单侧、双侧 *P* 值。右滚动条的属性规定 max 为 1000，min 为 0，linkedcell 为 J7，在 K6 单元格中键入公式“=J7*0.001”，在 J6 单元格中键入公式“=K6/2”，使滚动条在 0 至 0.5 之间变化，步幅为 0.0005。在单元格 L6 中键入公式“=TINV(K6,G2)”，获得双侧 α（单元格 K6）对应的 *t* 临界值。

（3）绘制 *t* 分布图

由 Excel 面积图第一个子图绘制 *t* 分布图（如图 19-7 左侧所示）。首先以单元格区域 A3:A803 为横坐标，B3:B803 为纵坐标，绘制整个 *t* 分布图；在同一图中，然后以单元格区域 A3:A803 为横坐标，C3:C803 为纵坐标，绘制以单元格 G2 为自由度，单元格 F6 对应 *t* 值以左的 *t* 分布图。图 19-7 右侧的 *t* 分布图绘制方法类似，只是由单元格区域 D3:D803 数据替代了 C3:C803。该图反映了 *t* 分布曲线下 *t* 值以右曲线下面积随单元格 G2、F6（由滚动条控制）的变化情况。

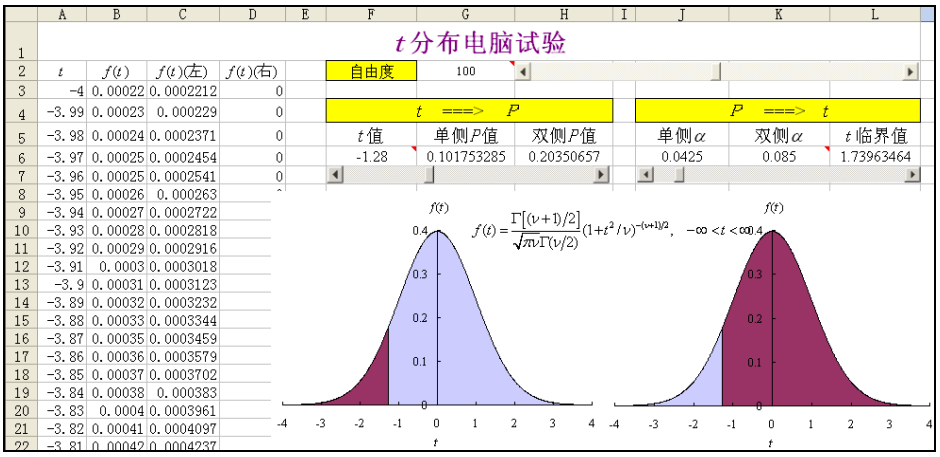


图 19-7　*t* 分布的电脑实验

电脑实验

打开光盘中的“19-3 t分布2.xlsx”文件，左右滑动自由度滚动条，可以发现：

①自由度越小，*t*分布中心位置纵高越矮，两侧尾部翘得越高。

②自由度30以下时，*t* 分布曲线形状改变明显，而当自由度较大时，*t*分布曲线形状改变不明显。

③自由度改变，*t*值不变时，单、双侧*P*值（单元格G6、H6）发生改变，自由度越小，单、双侧*P*值越大。

当左右滑动控制 *t* 值的滚动条时，左侧 *t* 分布图深色阴影左右移动，反映了 *t* 分布曲线下 *t* 值以左的曲线下面积随 *t* 值的变化；右侧 *t* 分布图深色阴影同时发生相应变化，反映了 *t* 分布曲线下 *t* 值以右的曲线下面积随 *t* 值的变化。

温馨提示：①为了与“由 *t* 值获得 *P* 值”形成对照，在该文件的单元格区域 J7:L7 处左右滑动右侧滚动条，可以“由 *P* 值（即 α 值）获得 *t* 临界值”。

②在“19-3 t 分布 2.xlsx”文件的下方，有一个图可同时修改、观看 *t* 分布曲线下 *t* 值以左与 *t* 值以右曲线下面积（如图 19-8 所示），其相应数据分别位于第 A、B、N、O 列。

③上面有些截图可能与光盘文件不一致，这是因为有关的参数（如自由度等）不相符的缘故，如果光盘文件相应参数修改为与截图相同的参数时，则光盘文件会与截图完全一致。

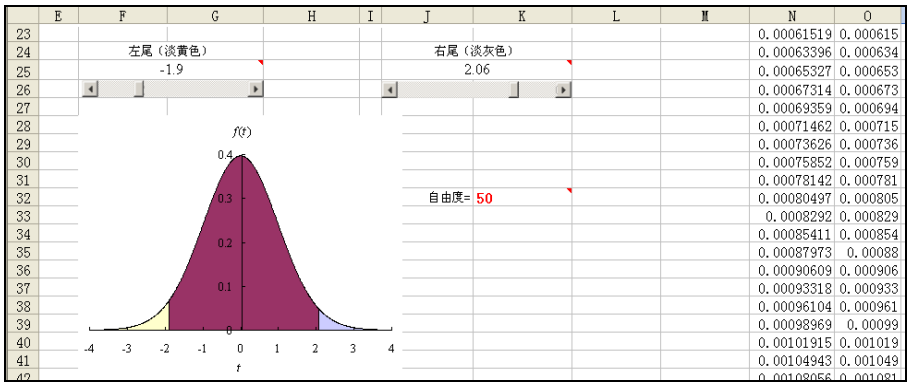


图 19-8 *t* 分布的双侧尾部面积的改变

19.3 *F* 分布

t 分布主要用于两样本均数的比较，当有多个均数需要比较时，需要采用 *F* 分布（*F* distribution）。*F* 分布的概率密度函数为：

$$f(F)=\frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)\nu_1^{\nu_1/2}\nu_2^{\nu_2/2}F^{\frac{\nu_1}{2}-1}}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)(\nu_1F+\nu_2)^{\frac{\nu_1+\nu_2}{2}}} \tag{19-12}$$

式中，Γ(•) 为伽玛函数；F=组间均方/组内均方=MS_1/MS_2，是两个均方的比值；ν_1、ν_2 分别为 F 值的分子、分母自由度，这是 F 分布的两个参数，由这两个自由度可决定 F 分布的图形形状，因此 F 分布可用 $F(\nu_1,\nu_2)$ 表示。以 F 为横轴，$f(F)$ 为纵轴可绘制 F 分布的图形。

19.3.1 F 分布的图形及其分布特征

例 19-8：在同一个图形中绘制 $F(1,5)$、$F(5,5)$、$F(10,10)$ 三个 F 分布图形。

解：先产生数据，再绘制图形。产生数据具体按如下步骤进行。

①在单元格区域 A1:D1 以及单元格区域 F3:H11 中，按图 19-9（见“19-4 F 分布 1.xlsx”文件）键入有关文字符号、数据等。

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	F值	F分布1	F分布2	F分布3		F分布1			F分布2			F分布3	
2	0.01	3.773381	0.012922	0.000006		自由度1(ν1)	1		自由度1(ν1)	5		自由度1(ν1)	10
3	0.05	1.647725	0.118973	0.002417		自由度2(ν2)	5		自由度2(ν2)	5		自由度2(ν2)	10
4	0.09	1.199414	0.238325	0.017460									
5	0.13	0.974810	0.345511	0.053007									
6	0.17	0.832814	0.434193	0.109466									
7	0.21	0.732185	0.503896	0.182123									
8	0.25	0.655837	0.556287	0.264241									
9	0.29	0.595220	0.593728	0.349169									
10	0.33	0.545513	0.618659	0.431400									
11	0.37	0.503756	0.633342	0.506923									
12	0.41	0.468015	0.639763	0.573172									
13	0.45	0.436966	0.639613	0.628797									
14	0.49	0.409665	0.634310	0.673383									
15	0.53	0.385417	0.625022	0.707180									
16	0.57	0.363698	0.612707	0.730876									
17	0.61	0.344105	0.598142	0.745420									
18	0.65	0.326319	0.581955	0.751890									
19	0.69	0.310086	0.564650	0.751398									
20	0.73	0.295202	0.546629	0.745025									
21	0.77	0.281495	0.528212	0.733788									
22	0.81	0.268827	0.509652	0.718613									
23	0.85	0.257079	0.491144	0.700324									

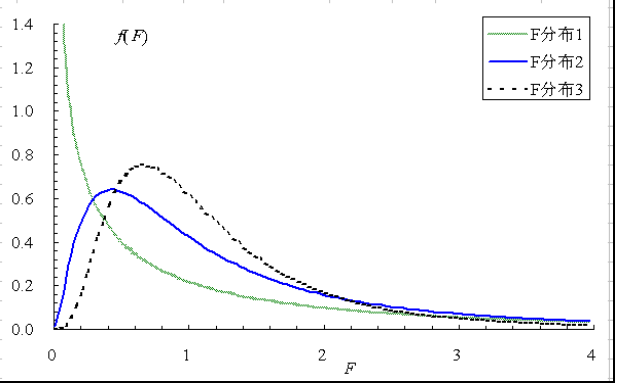


图 19-9　F 分布曲线

②在单元格 A2 中输入 0.01，用鼠标选中 A2，单击功能区的“开始→编辑→填充→系列”命令，在“序列”对话框中选择“序列产生在列”、“等差序列”，步长值、终止值分别填入 0.04、4，产生第 A 列 F 值的数据。

③为了产生随 F 值变化的 F 分布 1 的纵坐标值，按公式（19-12）在单元格 B2 中输入 Excel 公式：

“=EXP(GAMMALN(($G$2+$G$3)/2))*$G$2^($G$2/2)*$G$3^($G$3/2)*
$A2^($G$2/2-1)/EXP(GAMMALN($G$2/2))/EXP(GAMMALN($G$3/2))/
($G$2*$A2+$G$3)^(($G$2+$G$3)/2)”

为了产生 F 分布 2 的纵坐标值，按公式（19-12）在单元格 C2 中输入 Excel 公式：

“=EXP(GAMMALN(($J$2+$J$3)/2))*$J$2^($J$2/2)*$J$3^($J$3/2)*$A2^($J$2/2-1)/

EXP(GAMMALN($J$2/2))/EXP(GAMMALN($J$3/2))/($J$2*$A2+$J$3)^(($J$2+$J$3)/2)”

同样，为了产生 *F* 分布 3 的纵坐标值，按公式（19-12）在单元格 D2 中输入 Excel 公式：

“=EXP(GAMMALN(($M$2+$M$3)/2))*$M$2^($M$2/2)*$M$3^($M$3/2)
*$A2^($M$2/2-1)/EXP(GAMMALN($M$2/2))/EXP(GAMMALN($M$3/2))/
($M$2*$A2+$M$3)^(($M$2+$M$3)/2)”

④选取单元格区域 B2:D2，将此向下拖放至单元格 D101。

获得数据后，可利用这些数据及图表向导中的 *XY* 散点图，按下列步骤绘制图形。

①单击“插入→图表→散点图”图标的下拉窗口，选取子图表类型中的“带点平滑线的散点图”。

②在弹出的图表中，在空白处单击鼠标右键，在弹出的选项中单击“选择数据”命令，弹出“选择数据源”对话框，选择图表数据区域$A$1:$D$82，单击“确定”后即得到 *t* 分布图。

③为了使获得的图形美观好看，可以单击鼠标右键清除图域的背景色、网格线，改变横轴、纵轴刻度，添加横轴（*F*）、纵轴标题（*f*(*F*)），改变曲线颜色和样式，调整图例位置等。

④为了便于辨认，将存放各曲线名称、*F* 分布自由度的有关文字、数据颜色改变成与曲线一致。修改后的曲线图如图 19-9 所示。

电脑实验

反复修改 *F* 分布的自由度，观察比较不同分子、分母自由度的 *F* 分布曲线，体会 *F* 分布的如下性质：

①*F* 分布曲线随着自由度 v_1、v_1 的变化而变化。

②理论的 *F* 值取值范围应为 0 到无穷大，其平均值为 1（注意：*F* =1 两侧的概率不一定相等）。

③由于 *F* 值的平均值为 1，而在 *F* 值等于 1 处曲线不一定为最高点，所以 *F* 分布曲线一般为不对称分布。

19.3.2 *F* 分布曲线下面积与检验界值

可用 Excel 的 FDIST 函数计算 *F* 分布的右侧尾部曲线下面积。FDIST 函数的语法格式为“=FDIST(*F* 值，分子自由度，分母自由度)”。例如 *F* 值等于 3.74，分子、分母自由度分别为 2、14，由公式“=FDIST(3.74,2,14)”得检验 *P* 值为 0.05。

也可根据 *P* 值反过来求 *F* 界值，其函数的语法格式为“=FINV(*P* 值，分子自由度,分母自由度)”。例如 *P* 值等于 0.05，分子、分母自由度分别为 5、10，由公式“=FINV(0.05,5,10)”

得 F 界值为 3.3258；P 值等于 0.05，分子、分母自由度分别为 1、10，由公式“=FINV(0.05,1,10)”得 F 界值为 4.9646（见图 19-10）。

当分子自由度为 1 时，图 19-10 所给 F 界值表的 F 值开方等于 t 界值，即 $\sqrt{F}=t$。例如由“=TINV(0.05,10)”得到 α=0.05，自由度为 10 的 t 界值为 2.2281；P 值等于 0.05，分子、分母自由度分别为 1、10 的 F 界值为 4.9646，$\sqrt{4.9646}=2.2281$。

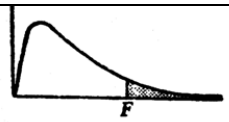
当 F 检验用于两样本方差齐性检验时，F 检验统计量的计算公式为：

$$F=\frac{S_1^2(\text{较大})}{S_2^2(\text{较小})},\quad \nu_1=n_1-1,\quad \nu_2=n_2-1 \qquad (19\text{-}13)$$

式中，S_1^2 为较大方差，S_2^2 为较小方差，相应的样本含量分别为 n_1 和 n_2，F 统计量服从分子自由度为 ν_1（第 1 自由度）、分母自由度为 ν_2（第 2 自由度）的 F 分布；求得 F 值后查方差齐性检验用表。该表与图 19-10 的 F 界值表略有不同，Excel 函数的语法格式为“=FINV(P/2 值，分子自由度，分母自由度)”。即方差检验用 $P=\alpha/2=0.05/2=0.025$ 的 F 界值（见图 19-10），相当于方差齐性检验用 $\alpha=0.05$ 的 F 界值。可能原因是：公式（19-13）只考虑了方差较大者作为分子，实际上也有可能方差较小者作为分子。例如，P 值等于 0.05，分子、分母自由度分别为 1、10，由公式“=FINV(0.05/2,1,10)”得方差齐性的 F 界值为 6.9367（如图 19-11 所示）。

附表 3　F 值表（方差分析检验用表）

上行：P=0.05 ，下行: P=0.01

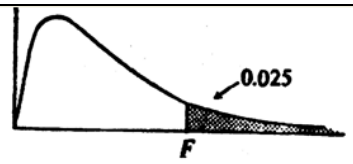


分母的自由度	分子的自由度，ν_1											
ν_2	1	2	3	4	5	6	7	8	9	10	11	12
1	161	200	216	225	230	234	237	239	241	242	243	224
	4052	4999	5403	5625	5764	5859	5928	5981	6022	6056	6082	6106
2	18.51	19.00	19.16	19.25	19.30	19.33	19.36	19.37	19.38	19.39	19.40	19.41
	98.49	99.00	99.17	99.25	99.30	99.33	99.34	99.36	99.38	99.40	99.41	99.42
3	10.13	9.55	9.28	9.12	9.01	8.94	8.88	8.84	8.81	8.78	8.76	8.74
	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.34	27.23	27.13	27.05
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.93	5.91
	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.54	14.45	14.37
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.78	4.74	4.70	4.68
	16.26	13.27	12.06	11.39	10.97	10.67	10.45	10.27	10.15	10.05	9.96	9.89
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00
	13.74	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.79	7.72
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.63	3.60	3.57
	12.25	9.55	8.45	7.85	7.46	7.19	7.00	6.84	6.71	6.62	6.54	6.47
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.34	3.31	3.28
	11.26	8.65	7.59	7.01	6.63	6.37	6.19	6.03	5.91	5.82	5.74	5.67
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.13	3.10	3.07
	10.56	8.02	6.99	6.42	6.06	5.80	5.62	5.47	5.35	5.26	5.18	5.11
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.97	2.94	2.91
	10.04	7.56	6.55	5.99	5.64	5.39	5.21	5.06	4.95	4.85	4.78	4.71

图 19-10　F 界值表（方差分析用）

附表 4　　F 值表（方差齐性检验用表）

P=0.05



分母的自由度	分子的自由度，ν_1															
ν_2	1	2	3	4	5	6	7	8	9	10	12	15	20	30	60	∞
1	647.79	799.50	864.16	899.58	921.85	937.11	948.22	956.66	963.29	968.63	976.71	984.87	993.10	1001.41	1009.80	1018.26
2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.41	39.43	39.45	39.46	39.4 8	39.50
3	17.44	16.04	15.44	15.10	14.88	14:73	14.62	14.54	14.47	14.42	14.34	14.25	14.17	14.08	13.99	13.90
4	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.75	8.66	8.56	8.46	8.36	8.26
5	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.52	6.43	6.33	6.23	6.12	6.02
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.37	5.27	5.17	5.07	4.96	4.85
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.67	4.57	4.47	4.36	4.25	4.14
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.20	4.10	4.00	3.89	3.78	3.67
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.87	3.77	3.67	3.56	3.45	3.33
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.62	3.52	3.42	3.31	3.20	3.08
11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	3.43	3.33	3.23	3.12	3.00	2.88
12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.28	3.18	3.07	2.96	2.85	2.72
13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	3.15	3.05	2.95	2.84	2.72	2.60
14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	3.05	2.95	2.84	2.73	2.61	2.49
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.96	2.86	2.76	2.64	2.52	2.40
16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99	2.89	2.79	2.68	2.57	2.45	2.32
17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92	2.82	2.72	2.62	2.50	2.38	2.25
18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87	2.77	2.67	2.56	2.44	2.32	2.19
19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82	2.72	2.62	2.51	2.39	2.27	2.13
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.68	2.57	2.46	2.35	2.22	2.09
21	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80	2.73	2.64	2.53	2.42	2.31	2.18	2.04
22	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70	2.60	2.50	2.39	2.27	2.14	2.00
23	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73	2.67	2.57	2.47	2.36	2.24	2.11	1.97
24	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64	2.54	2.44	2.33	2.21	2.08	1.94
25	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61	2.51	2.41	2.30	2.18	2.05	1.91
26	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65	2.59	2.49	2.39	2.28	2.16	2.03	1.88
27	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63	2.57	2.47	2.36	2.25	2.13	2.00	1.85
28	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61	2.55	2.45	2.34	2.23	2.11	1.98	1.83
29	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59	2.53	2.43	2.32	2.21	2.09	1.96	1.81

图 19-11　F 界值表（方差齐性检验用）

19.3.3　F 分布电脑实验

F 分布电脑实验存放在光盘文件“19-5 F 分布 2.xlsx”中，数据的准备、滚动条的制作、F 分布图的绘制与 z 分布、t 分布类似，在此不再赘述。

电脑实验

打开光盘中的“19-5 F分布2.xlsx”文件（如图19-12所示），左右滑动分子、分母自由度滚动条，可以发现F分布曲线形状改变明显。

当左右滑动控制F 值的滚动条时，F分布右侧深色阴影左右移动，反映了F分布曲线下F值以右的曲线下面积随F值的变化。

当左右滑动控制F值的滚动条时，对应于单元格E6中F值的P值（单元格F6）发生相应改变。为了与“由F值获得P值”形成对照，在该文件的单元格区域I7:K7处左右滑动右侧滚动条，可以“由P值（即α值）获得F临界值”。

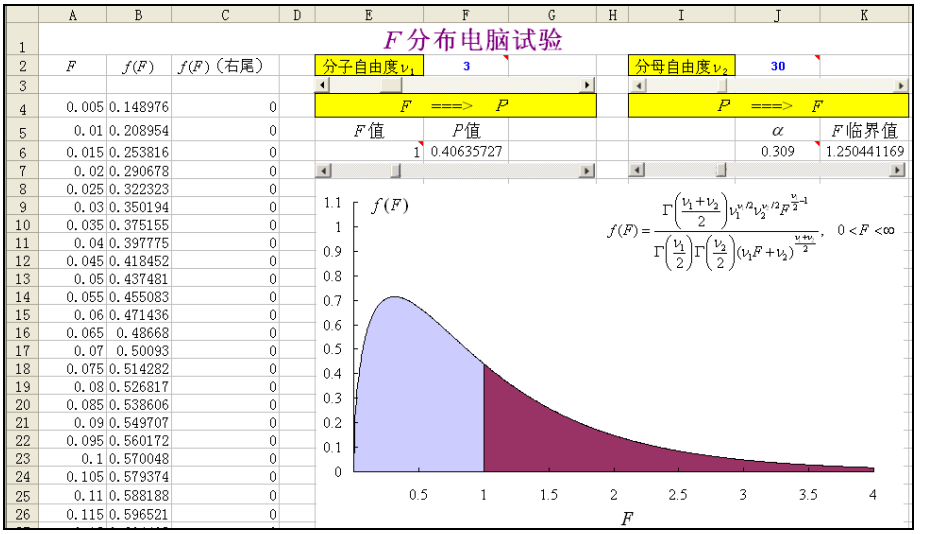


图 19-12　F 分布的电脑实验

19.4　χ^2 分布

尽管 χ^2 分布（Chi-square distribution）常用于离散资料的检验，但它属于连续分布。χ^2 变量是相互独立的标准正态变量 z 的平方和，即

$$\chi^2 = z_1^2 + z_2^2 + \cdots + z_\nu^2$$

χ^2 变量服从自由度为 ν 的 χ^2 分布。χ^2 分布的概率密度函数为：

$$f(\chi^2) = \frac{1}{2\Gamma(\nu/2)}\left(\frac{\chi^2}{2}\right)^{(\nu/2-1)} \mathrm{e}^{-\chi^2/2} \qquad (19\text{-}14)$$

式中，$\Gamma(\bullet)$ 为伽玛函数；e 为自然对数的底，其值为 2.71828…；ν 为自由度，是 χ^2 分布的唯一参数，由此决定 χ^2 分布的图形形状，因此 χ^2 分布可用 $\chi^2(\nu)$ 表示。以 χ^2 为横轴，$f(\chi^2)$ 为纵轴可绘制 χ^2 分布的图形。

19.4.1　χ^2 分布的图形及其分布特征

例 19-9：在同一张图中绘制 $\chi^2(1)$、$\chi^2(2)$、$\chi^2(3)$、$\chi^2(6)$ 四条 χ^2 分布曲线。

解：先产生数据，再绘制图形。产生数据具体按如下步骤进行。

①在单元格区域 A1:E1 以及单元格区域 G1:N2 中，按图 19-13（见光盘文件“19-6 卡方分布 1.xlsx”）键入有关文字符号、数据等。

	A	B	C	D	E	G	H	I	J	K	L	M	N
1	卡方值	卡方分布1	卡方分布2	卡方分布3	卡方分布4	卡方分布1		卡方分布2		卡方分布3		卡方分布4	
2	0.0					自由度	1	自由度	2	自由度	3	自由度	6
3	0.2	0.807171	0.452419	0.161434	0.002262								
4	0.4	0.516442	0.409365	0.206577	0.0								
5	0.6	0.381545	0.370409	0.228927	0.0								
6	0.8	0.298984	0.335160	0.239187	0.0								
7	1.0	0.241971	0.303265	0.241971	0.0								
8	1.2	0.199868	0.274406	0.239841	0.0								
9	1.4	0.167433	0.248293	0.234406	0.0								
10	1.6	0.141715	0.224664	0.226743	0.0								
11	1.8	0.120895	0.203285	0.217611	0.0								
12	2.0	0.103777	0.183940	0.207554	0.0								
13	2.2	0.089531	0.166436	0.196969	0.1								
14	2.4	0.077562	0.150597	0.186150	0.1								
15	2.6	0.067428	0.136266	0.175313	0.1								
16	2.8	0.058792	0.123298	0.164618	0.1								
17	3.0	0.051393	0.111565	0.154180	0.1								
18	3.2	0.045026	0.100948	0.144083	0.1								
19	3.4	0.039525	0.091342	0.134384	0.1								
20	3.6	0.034756	0.082649	0.125121	0.1								
21	3.8	0.030610	0.074784	0.116317	0.1								
22	4.0	0.026995	0.067668	0.107982	0.1								
23	4.2	0.023838	0.061228	0.100119	0.1								
24	4.4	0.021073	0.055402	0.092723	0.1								
25	4.6	0.018649	0.050129	0.085785	0.1								

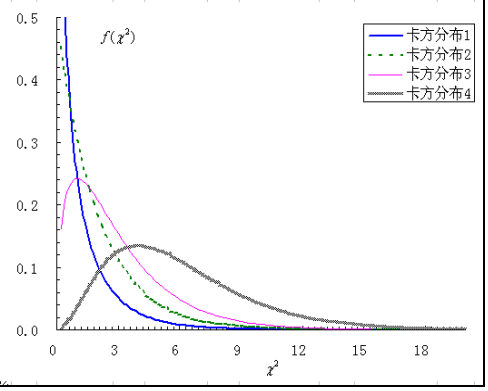


图 19-13 χ^2 分布曲线

②在单元格 A2 中输入 0，单元格 A3 中输入 0.2，用鼠标选中单元格区域 A2:A3，向下拖放填充至单元格 A102。

③为了产生随卡方（χ^2）值变化的卡方分布 1、2、3、4 的纵坐标值，按公式（19-14）分别在单元格 B3、C3、D3、E3 中输入 Excel 公式：

“=1/(2*EXP(GAMMALN($H$2/2)))*($A3/2)^($H$2/2-1)*EXP(($A3/2)*(-1))”

“=1/(2*EXP(GAMMALN($J$2/2)))*($A3/2)^($J$2/2-1)*EXP(($A3/2)*(-1))”

“=1/(2*EXP(GAMMALN($L$2/2)))*($A3/2)^($L$2/2-1)*EXP(($A3/2)*(-1))”

“=1/(2*EXP(GAMMALN($N$2/2)))*($A3/2)^($N$2/2-1)*EXP(($A3/2)*(-1))”

④将单元格区域 B3:E3 中的公式复制、粘贴到单元格区域 B4: E102。

获得数据后，按图 19-13 适当调整小数位数。利用这些数据及图表向导中的 *XY* 散点图，按下列步骤绘制图形。

①单击“插入→图表→散点图”图标的下拉窗口，选取子图表类型中的“带平滑线的散点图”。

②在弹出的图表中，在空白处单击鼠标右键，在弹出的选项中单击“选择数据”命令，弹出“选择数据源”对话框，选择图表数据区域$A$1:$E$102，单击“确定”按钮后即得到 χ^2 分布图。

③为了使获得的图形美观好看，可以单击鼠标右键清除图域的背景色、网格线，改变横轴、纵轴刻度，添加横轴（纵高）、纵轴（卡方值）标题，改变曲线颜色和样式，调整图

例位置等。

④为了便于辨认，将存放各曲线名称、χ^2 分布自由度的有关文字、数据颜色改变成与曲线一致。修改后的曲线图如图 19-13 所示。

电脑实验

反复修改“19-6 卡方分布 1.xlsx”文件中某 χ^2 分布曲线的自由度（如修改单元格 H2），观察比较不同自由度的 χ^2 分布曲线，体会 χ^2 分布的如下性质：

①χ^2 分布曲线形状随着自由度 v 的变化而变化，不同自由度下 χ^2 分布曲线不同。

②χ^2 分布为连续分布，其理论的 χ^2 取值范围应为 0 到无穷大。

③χ^2 分布曲线左右不对称，峰值偏向于左侧，右侧尾部曲线拖尾（即右偏态）。自由度较小时偏度较大，随着自由度的增大，偏度减少。

19.4.2 χ^2 分布曲线下面积与检验界值

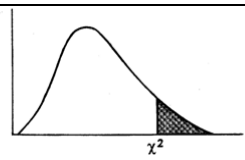
可用 Excel 的 CHIDIST 函数计算 χ^2 分布的右侧尾部的曲线下面积。CHIDIST 函数的语法格式为“=CHIDIST(χ^2 值，自由度)”。例如 χ^2 值等于 6.63，自由度为 1，由公式“=CHIDIST(6.63,1)”得检验 P 值为 0.0100。目前 CHIDIST 改变成了 CHISQ.DIST 以及 CHISQ.DIST.RT 几个函数，其中“=CHISQ.DIST.RT(6.33,1)”与“=CHIDIST(6.63,1)”完全等价；而 CHISQ.DIST 函数多了一个参数，该参数为“1”时获得累积分布函数，为“0”时获得概率密度函数。

也可根据 P 值反过来求 χ^2 界值，其函数的语法格式为“=CHIINV(P 值，自由度)”。例如 P 值等于 0.01，自由度分别为 1，由公式“=CHIINV(0.01,1)”得 χ^2 界值为 6.6349；例如 P 值等于 0.05，自由度分别为 15，由公式“=CHIINV(0.05,15)”得 χ^2 界值为 24.9958（见图 19-14）。目前 CHIINV 改变成了 CHISQ.INV 以及 CHISQ.INV.RT 几个函数，其中“=CHISQ.INV.RT(0.05,15)”与“=CHIINV(0.05,15)”完全等价，都是返回右尾概率对应的临界值；而 CHISQ.INV 函数返回的是左尾概率对应的临界值。

19.4.3 χ^2 分布电脑实验

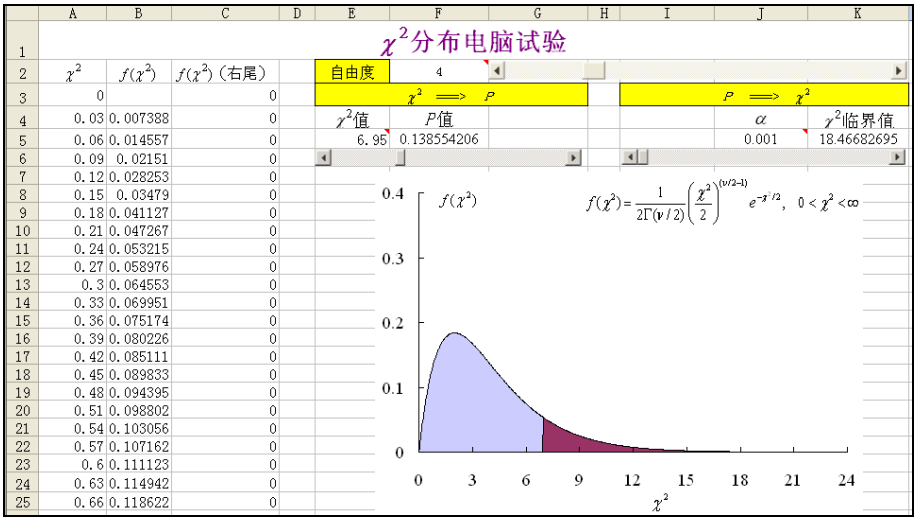
χ^2 分布电脑实验存放在光盘文件“19-7 卡方分布 2.xlsx”中，数据的准备、滚动条的制作、χ^2 分布图的绘制与 z 分布、t 分布类似，在此不再赘述。

附表 5　χ^2 界值表



自由度	概率，P												
ν	0.995	0.990	0.975	0.950	0.900	0.750	0.500	0.250	0.100	0.050	0.025	0.010	0.005
1					0.02	0.10	0.45	1.32	2.71	3.84	5.02	6.63	7.88
2	0.01	0.02	0.05	0.10	0.21	0.58	1.39	2.77	4.61	5.99	7.38	9.21	10.60
3	0.07	0.11	0.22	0.35	0.58	1.21	2.37	4.11	6.25	7.81	9.35	11.34	12.84
4	0.21	0.30	0.48	0.71	1.06	1.92	3.36	5.39	7.78	9.49	11.14	13.28	14.86
5	0.41	0.55	0.83	1.15	1.61	2.67	4.35	6.63	9.24	11.07	12.83	15.09	16.75
6	0.68	0.87	1.24	1.64	2.20	3.45	5.35	7.84	10.64	12.59	14.45	16.81	18.55
7	0.99	1.24	1.69	2.17	2.83	4.25	6.35	9.04	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	5.07	7.34	10.22	13.36	15.51	17.53	20.09	21.95
9	1.73	2.09	2.70	3.33	4.17	5.90	8.34	11.39	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	6.74	9.34	12.55	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	7.58	10.34	13.70	17.28	19.68	21.92	24.72	26.76
12	3.07	3.57	4.40	5.23	6.30	8.44	11.34	14.85	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	9.30	12.34	15.98	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	10.17	13.34	17.12	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.26	7.26	8.55	11.04	14.34	18.25	22.31	25.00	27.49	30.58	32.80

图 19-14　χ^2 界值表



χ^2	$f(\chi^2)$	$f(\chi^2)$（右尾）
0		0
0.03	0.007388	0
0.06	0.014557	0
0.09	0.02151	0
0.12	0.028253	0
0.15	0.03479	0
0.18	0.041127	0
0.21	0.047267	0
0.24	0.053215	0
0.27	0.058976	0
0.3	0.064553	0
0.33	0.069951	0
0.36	0.075174	0
0.39	0.080226	0
0.42	0.085111	0
0.45	0.089833	0
0.48	0.094395	0
0.51	0.098802	0
0.54	0.103056	0
0.57	0.107162	0
0.6	0.111123	0
0.63	0.114942	0
0.66	0.118622	0

图 19-15　χ^2 分布的电脑实验

电脑实验

打开光盘文件“19-7 卡方分布 2.xlsx”（如图 19-15 所示），左右滑动自由度滚动条，可

以发现 χ^2 分布曲线形状改变明显。χ^2 分布曲线左右不对称，峰值偏向于左侧，极端值向右侧拖尾（即右偏态）。自由度较小时偏度较大，随着自由度的增大，偏度逐步减少。χ^2 分布的极限分布为正态分布。

当左右滑动控制 χ^2 值的滚动条时，χ^2 分布右侧深色阴影左右移动，反映了 χ^2 分布曲线下 χ^2 值以右的曲线下面积随 χ^2 值的变化。

当左右滑动控制 χ^2 值的滚动条时，对应于单元格 E5 中 χ^2 值的 P 值（单元格 F5）发生相应改变。为了与“由 χ^2 值获得 P 值”形成对照，在该文件的单元格区域 I6:K6 处左右滑动滚动条，可以“由 P 值（即 α 值）获得 χ^2 临界值”。

本章所提供的 z 分布、t 分布、F 分布、χ^2 分布电脑实验可直观、动态地反映这些分布的分布特征。此外，通过滑动滚动条，不仅可精确获得由检验统计量推算的具体 P 值，而且可通过检验水准 α 等参数精确获得检验统计量临界值。

第20章

其他电脑实验与置信区间估计

随着电脑的普及，电脑实验在统计学中的地位越来越明显。前面已介绍了各种概率分布、随机抽样等电脑实验，本章将在此基础上根据统计学应用中的实际问题给出若干个其他电脑实验，起抛砖引玉的作用，但愿以后有更多类似电脑实验供学习者或研究者参考。

20.1 随机现象

统计学是研究大量同类随机现象的统计学规律性的学科，不研究随机性、不研究变异就谈不上是真正的统计学。

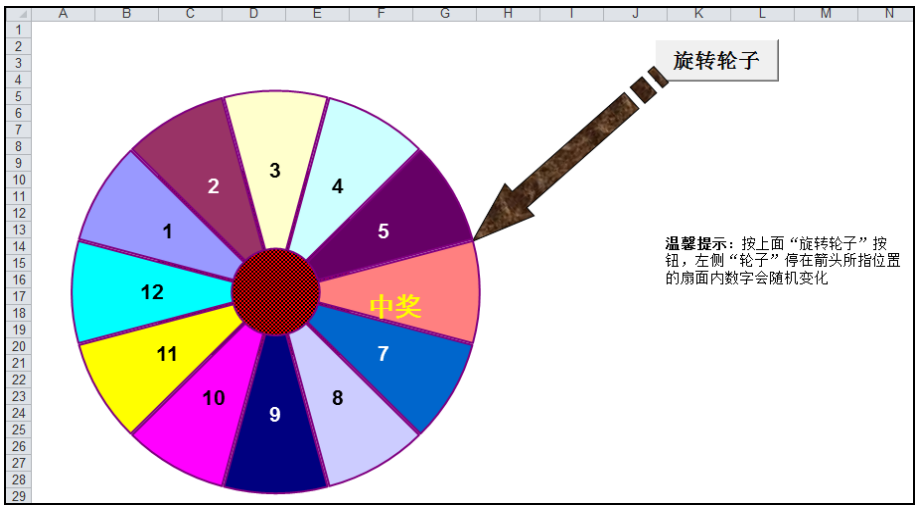


图 20-1　随机现象

光盘文件“20-1 随机现象.xlsm（如图 20-1 所示），是根据 Excel 宏编制的程序（参见附录 C），其宏程序为：

```
Option Explicit
  Sub SpinTheWheel()
     Dim Cg AsChartGroup
     Dim knt As Long, n As Long, a As Long, Inc As Long
knt = 0
     Randomize
     Set Cg = ActiveSheet.ChartObjects(1).Chart.ChartGroups(1)
    a = Rnd() * 360
Cg.FirstSliceAngle = a
Inc = 20 + Int(Rnd() * 4)
     Do
knt = knt + 1
        a = (a + Inc) Mod 360
Cg.FirstSliceAngle = a
DoEvents
        If Inc< 2 Then
            If Rnd() < 0.05 Then
                Exit Sub
            End If
        End If
DoEvents
knt = knt + 1
```

```
        If knt Mod 20 = 0 Then If Inc<> 1 Then Inc = Inc - 1
    Loop
End Sub
```

图 20-1 左侧的轮子通过 Excel 圆环图制作而成，分类轴为 1,2,……,12；值为 12 个 1，其中“6”有“中奖”两字替代。

电脑实验

打开光盘文件“20-1 随机现象.xlsm”，单击上面的“旋转轮子”按钮，左侧“轮子”会自动旋转 10 秒钟左右（具体时间与计算机性能有关），停在箭头所指位置的扇面内数字会随机变化，每次旋转结果均有可能不相同。

温馨提示：因为文件中包括 Excel 宏程序，所以应该设置宏安全性。方法是选择“开发工具→代码→宏安全性”，在弹出的“信任中心”对话框中，单击“宏设置”，将宏安全性改为“禁用无数字签署的所有宏”，然后单击“确定”即可。再使用完宏后应将宏安全性重新设置为“禁用所有宏”，以保证计算机安全。

对于未设置宏安全性的用户，打开此类 Excel 文件时，会在功能区下方出现“安全警告宏已被禁用”，单击安全警告后的“启用内容”选项，即可正常运行。

20.2 置信区间

参数估计有点估计和区间估计两种。用样本统计量直接作为总体参数的估计值就是点估计，如直接用随机样本的样本均数 $\bar{X}$ 作为总体均数 μ 的点估计值。点估计方法简单，但未考虑抽样误差的影响，估计的正确程度很难评价。

区间估计是按事先给定的概率 $(1-\alpha)$，估计包含未知总体参数的一个区间范围，该范围称为参数的置信区间（Confidence Interval，CI），其估计方法称区间估计。$(1-\alpha)$ 称为置信度（Confidence Level）或置信系数（Confidence Coefficient），也可表示为 $100(1-\alpha)\%$，常取 95%（也取 90%、99%）。置信区间通常由两个数值即两个置信限（Confidence Limit，CL）表示，较小者称为置信下限（Lower limit，L），较大者称为置信上限（Upper limit，U）。

置信区间有准确度（accuracy）与精确度（precision）两个要素。准确度由置信度$(1-\alpha)$的大小确定，即由“此区间包含总体参数”这句话可信程度$(1-\alpha)$的大小来反映，从准确度的角度看，置信度愈接近于 1 愈好，如置信度 99%比 95%好；精确度是置信区间宽度的一半（即 $t_{\alpha/2,\nu}S_{\bar{X}}$、$z_{\alpha/2,\nu}S_p$），意指置信区间的两端点值离样本统计量（如 $\bar{X}$、p）的距离。从精确度的角度看，置信区间宽度愈窄愈好。在抽样误差确定的情况下，二者是相互

矛盾的。若提高了置信度，即α减小，则检验统计量界值（如$t_{\alpha/2,\nu}$、$z_{\alpha/2}$）增大，置信区间变宽，从而导致精确度下降；反之，降低置信度，即降低准确度，可适当增加置信区间的精确度。为了同时兼顾置信区间的准确度与精确度，可适当增加样本含量；在置信度确定的情况下，增加样本含量可降低抽样误差，从而缩小置信区间范围，提高参数估计的精确度。

20.2.1 置信区间的电脑实验

置信区间的确切含义是指：如果能够进行重复抽样试验，则平均有$1-\alpha$（如 95%）的置信区间包含了总体参数。例如，总体均数的 95%置信区间的实际含义是：如果从同一总体中重复抽取 100 份样本含量相同的独立样本，每份样本分别计算 1 个置信区间，在 100 个置信区间中，将大约有 95 个置信区间覆盖总体均数，大约有 5 个置信区间并不覆盖总体均数。对于某一次估计的置信区间，根据小概率事件不太可能在一次试验中发生的原理，我们总是宣称这个区间覆盖了总体均数，但不一定是真的覆盖了总体均数，于是，我们补充一句：置信度为 95%，该结论错误的概率小于或等于 0.05。

图 20-2（对应于光盘文件“20-2 置信区间.xlsx”）的第 B 至 F 列共 5 列数据是满足正态分布 $N(5,1^2)$ 的随机样本，即在 B3 单元格中键入“=NORMINV(RAND(), $G$2, $H$2)”，将此公式向下拖放至 F102，产生样本含量 n=5 的 100 份样本。G2、H2 为正态总体参数，用户可以修改。单元格区域 G3:G102 与 H3:H102 分别为 100 份样本的每 5 个个体值计算获得的样本均数与样本标准差。J3:J102 与 K3:K102 分别为 100 份样本获得的 95%置信区间的下限值与上限值，如单元格 J3 为“=$G3−TINV(1−$J$1,4)*$H3/(5^0.5)”，单元格 K3 为“=$G3+TINV(1−$J$1,4)*$H3/(5^0.5)”。用户可以在单元格 J1 中修改置信度。单元格 N4 的公式为“=(COUNTIF(J3:J102,">"&G2)+COUNTIF(K3:K102,"<"&G2))/100”，计算了置信区间不覆盖总体均数的百分比。



电脑实验

打开光盘文件“20-2 置信区间.xlsx”，可进行如下操作：

①反复按键盘上的 F9 键（产生新的随机数据），观察单元格 N4 中的百分比的变化，并与单元格 J1 中的置信度比较。一般情况下，不覆盖总体均数的百分比通常在α（如 5%=0.05）附近上下波动。

②修改 J1 中的置信度，按 F9 键，观察 N4 单元格中的百分比变化。单元格 N4 中百分比会随 J1 增大而减少。

③修改单元格 G2、H2 正态总体参数，按 F9 键，观察 N4 单元格中的百分比变化。

④观察文件中的置信度图形，可直观查看哪些样本的置信区间不包括总体均数。该图主要由 J3：K102 单元格区域的 100 个置信区间组成，每一条竖线代表一个置信区间。中间横线表示总体均数大小，置信区间跨越该横线表示包括总体均数，否则不包括。哪些置信区间“不包括”总体均数，也可查看 J3：K102 单元格区域，红色字体对应的置信区间表示不包括总体均数。每按一次“F9”按钮，置信区间图形都会同步改变。

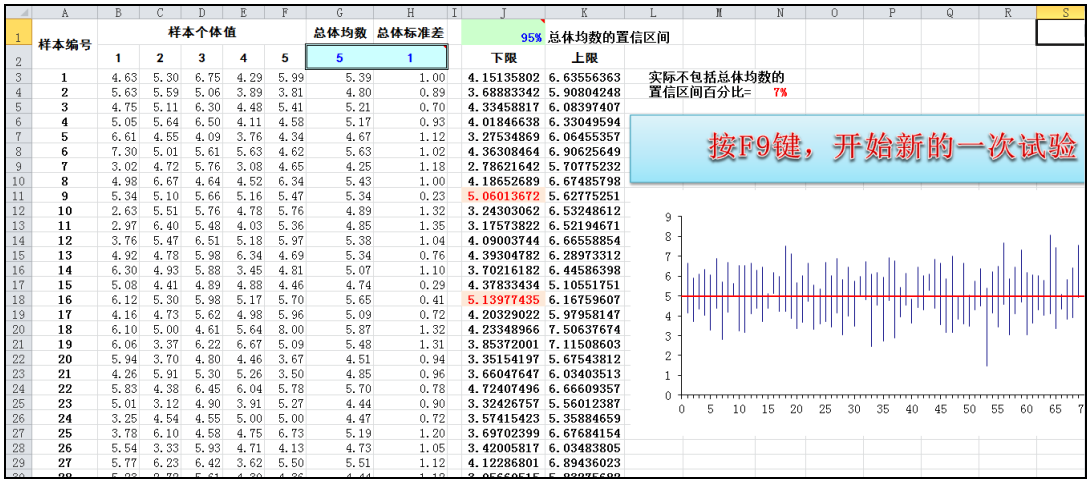


图 20-2 置信区间的电脑实验

20.2.2 置信区间的估计

区间估计与假设检验（参见第 8、9 章），是由样本数据对总体参数作出统计学推断的两种主要方法。置信区间用于说明量的大小，即推断总体参数的置信范围；而假设检验用于推断质的不同，即判断两个总体参数是否不等。两者既相互联系，又有所区别。

一方面，置信区间亦可回答假设检验的问题，若算得的置信区间包含了总体参数值，则按 α 检验水准，不拒绝 H_0；若不包含总体参数值，则按 α 水准，拒绝 H_0，接受 H_1。

另一方面，置信区间的宽窄反映了数据的精确度大小。如果计算得到的置信区间较宽，说明数据的精确度不高，对于相同的研究主题，精确度不高很可能是由于样本含量不够所引起的。

虽然置信区间亦可回答假设检验的问题，并能提供更多的信息，但并不意味着置信区间能够完全代替假设检验。置信区间只能在预先规定的 I 型错误概率，即检验水准为 α 的前提下进行计算，而假设检验能够获得一个较为确切的概率 P 值。故一般推荐将置信区间的估计与假设检验的推断相互结合使用，达到取长补短之目的。

在光盘文件“20-3 置信区间的估计.xlsx”中包含了方差已知（或大样本含量）总体均数的置信区间估计、方差未知（或小样本含量）总体均数的置信区间估计、总体率的置信区间估计、总体方差的置信区间估计共计 4 个工作表。

1. 方差已知（或大样本含量）总体均数的置信区间估计

在总体方差 σ^2 已知，或样本含量很大（如 $n \geqslant 100$）的情况下，进行均数比较时可以采用。计算公式已经存放在有关单元格附近，$z_{\alpha/2}$ 为标准正态分布曲线下两侧尾部面积各占 $\alpha/2$ 的右侧临界值（正数），z_{α} 是标准正态分布曲线下右侧尾部面积为 α 的界值，$\sigma_{\bar{X}} = \sigma/\sqrt{n}$，$n$ 为样本含量。

电脑实验

打开光盘文件“20-3 置信区间的估计.xlsx”的“1. 方差已知均数比较”工作表（见图 20-3），在“请输入已知数据”右侧键入数据，其下的“结果”会自动修正。

均数的置信区间（Confidence Interval，CI）估计

（1） 方差已知的总体均数CI　　置信度为95%

	α	n	$\bar{X}$	σ
请输入已知数据	0.05	100	27.191	8

结果

		CI宽度的一半	下限（L）	上限（U）	
双侧		1.567971188	25.62302881	28.7589712	$\left(\bar{X}-z_{\alpha/2}\sigma_{\bar{X}},\ \bar{X}+z_{\alpha/2}\sigma_{\bar{X}}\right)$
单侧	总体参数位于右尾	1.315882902	25.8751171	∞	$(\bar{X}-z_{\alpha}\sigma_{\bar{X}},\ +\infty)$
	总体参数位于左尾	1.315882902	－∞	28.5068829	$(-\infty,\ \bar{X}+z_{\alpha}\sigma_{\bar{X}})$

（2） 方差已知两总体均数之差的CI　　置信度为95%

	α	n_1	$\bar{X}_1$	σ_1	n_2	$\bar{X}_2$	σ_2
请输入已知数据	0.05	10	3	1	11	2	1

结果

		CI宽度的一半	下限（L）	上限（U）	
双侧		0.856369903	0.143630097	1.8563699	$\left(\lvert\bar{X}_1-\bar{X}_2\rvert-z_{\alpha/2}\sigma_{\bar{X}_1-\bar{X}_2},\ \lvert\bar{X}_1-\bar{X}_2\rvert+z_{\alpha/2}\sigma_{\bar{X}_1-\bar{X}_2}\right)$
					$\sigma_{\bar{X}_1-\bar{X}_2}=\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}$
单侧	总体参数位于右尾	0.718688278	0.281311722	∞	$\left(\lvert\bar{X}_1-\bar{X}_2\rvert-z_{\alpha}\sigma_{\bar{X}_1-\bar{X}_2},\ +\infty\right)$
	总体参数位于左尾	0.718688278	－∞	1.71868828	$\left(-\infty,\ \lvert\bar{X}_1-\bar{X}_2\rvert+z_{\alpha}\sigma_{\bar{X}_1-\bar{X}_2}\right)$

图 20-3　大样本或方差已知的总体均数置信区间

实例与结果：如果已知一组大样本，样本含量 $n=100$，均数 $\bar{X}=27.191$，标准差 $S=8$，则总体均数 95%置信区间为（25.62,28.76），反映精确度的置信区间宽度的一半为 1.57。如果研究需要获得单侧置信区间，该程序也已经给出。

温馨提示：该程序提供了获得单个总体均数和两个总体均数之差的置信区间估计程序，可根据需要选用。该程序适合方差已知或样本含量很大的总体均数比较。

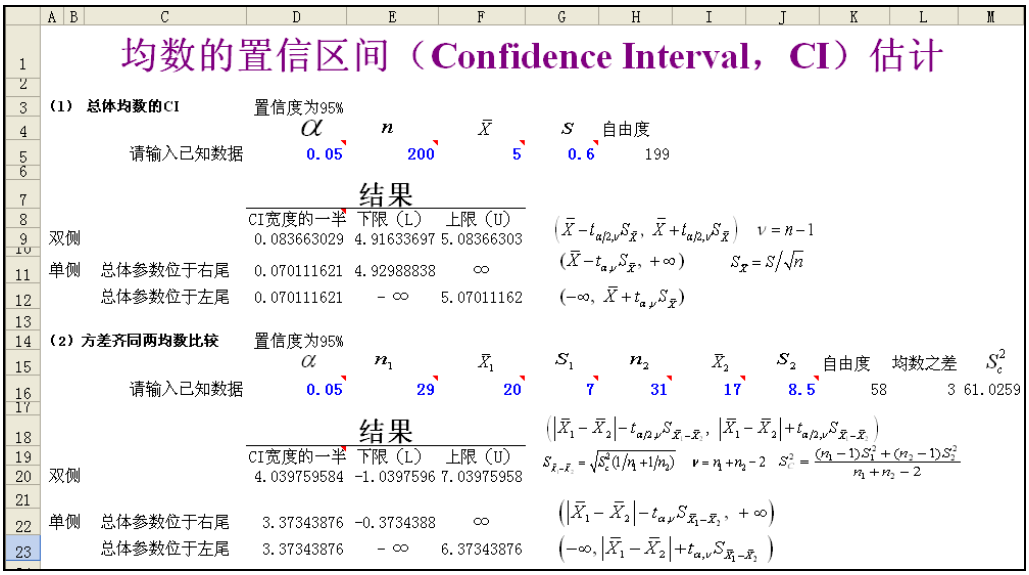
2. 方差未知（或小样本含量）总体均数的置信区间估计

在总体方差σ^2未知，或样本含量较小（如$n<30$）的情况下，进行均数比较时可以采用。计算公式已经存放在有关单元格附近，$t_{\alpha/2,\nu}$与$t_{\alpha,\nu}$分别为双侧与单侧α的界值，其他同上。

电脑实验

打开光盘文件“20-3 置信区间的估计.xlsx”的“2. 方差未知均数比较”工作表（如图20-4所示），在“请输入已知数据”右侧键入数据，其下的“结果”会自动修正。

实例与结果：例如，为了解氨甲喋呤对外周血IL-2水平的影响，某医生将60名哮喘患者随机分为两组。其中对照组29例，采用安慰剂；实验组31例，采用小剂量氨甲喋呤进行治疗。测得对照组治疗前IL-2的均数为20.00 IU/ml，标准差为7.00 IU/ml；试验组治疗前IL-2的均数为17.00 IU/ml ，标准差为8.50 IU/ml，请计算两个总体均数差值的置信区间。该数据属于“方差齐同两均数比较”，所以在程序的第16行键入已知数据，得到两个总体均数之差的95%置信区间为（–1.04, 7.04），反映精确度的置信区间宽度的一半为4.04。如果研究需要获得单侧置信区间，该程序也已经给出。



均数的置信区间（Confidence Interval，CI）估计

（1） 总体均数的CI　置信度为95%

	α	n	$\bar{X}$	s	自由度
请输入已知数据	0.05	200	5	0.6	199

结果

		CI宽度的一半	下限（L）	上限（U）	
双侧		0.083663029	4.91633697	5.08366303	$(\bar{X}-t_{\alpha/2,\nu}S_{\bar{X}},\ \bar{X}+t_{\alpha/2,\nu}S_{\bar{X}})$　$\nu=n-1$
单侧	总体参数位于右尾	0.070111621	4.92988838	∞	$(\bar{X}-t_{\alpha,\nu}S_{\bar{X}},\ +\infty)$　$S_{\bar{X}}=S/\sqrt{n}$
	总体参数位于左尾	0.070111621	- ∞	5.07011162	$(-\infty,\ \bar{X}+t_{\alpha,\nu}S_{\bar{X}})$

（2） 方差齐同两均数比较　置信度为95%

	α	n_1	$\bar{X}_1$	S_1	n_2	$\bar{X}_2$	S_2	自由度	均数之差	S_c^2
请输入已知数据	0.05	29	20	7	31	17	8.5	58	3	61.0259

结果

		CI宽度的一半	下限（L）	上限（U）	
双侧		4.039759584	-1.0397596	7.03975958	$\left(\lvert\bar{X}_1-\bar{X}_2\rvert-t_{\alpha/2,\nu}S_{\bar{X}_1-\bar{X}_2},\ \lvert\bar{X}_1-\bar{X}_2\rvert+t_{\alpha/2,\nu}S_{\bar{X}_1-\bar{X}_2}\right)$　$S_{\bar{X}_1-\bar{X}_2}=\sqrt{S_c^2(1/n_1+1/n_2)}$　$\nu=n_1+n_2-2$　$S_c^2=\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$
单侧	总体参数位于右尾	3.37343876	-0.3734388	∞	$\left(\lvert\bar{X}_1-\bar{X}_2\rvert-t_{\alpha,\nu}S_{\bar{X}_1-\bar{X}_2},\ +\infty\right)$
	总体参数位于左尾	3.37343876	- ∞	6.37343876	$\left(-\infty,\ \lvert\bar{X}_1-\bar{X}_2\rvert+t_{\alpha,\nu}S_{\bar{X}_1-\bar{X}_2}\right)$

图20-4　t检验对应的置信区间

温馨提示：该程序提供了获得单个总体均数、方差齐同两个总体均数之差、方差不齐两个总体均数之差（图20-4中没有显示）的置信区间估计共计3个程序，可根据需要选用。

该程序适合方差已知或样本含量很大的总体均数比较。样本均数与总体均数的比较，以及配对 t 检验资料，均可采用“（1）总体均数的 CI”，以配对 t 检验的两个样本差值的均数 $\bar{d}$ 替代 $\bar{X}$，差值的标准差 S_d 替代 S，对子数替代 n 即可；两个样本均数比较的置信区间采用同一工作表中的后两个程序。

3. 总体率的置信区间估计

当样本含量很小时，可直接利用二项分布原理获得其置信区间；当样本含量较大时，可利用近似正态分布原理获得率的置信区间。大样本计算置信区间的计算公式已经存放在有关单元格附近，p、p_1、p_2 表示样本率，S_p 与 $S_{p_1-p_2}$ 分别表示样本率 p 与样本率之差 p_1-p_2 所对应的率的标准误，其他同上。

电脑实验

打开光盘文件“20-3 置信区间的估计.xlxs”的“3. 率的置信区间”工作表（如图 20-5 所示），在“请输入已知数据”右侧键入数据，其下的“结果”会自动修正。

率的置信区间（Confidence Interval，CI）估计

（1） 单个总体率的CI

		置信度为95% α	n	p	S_p
请输入已知数据		0.05	250	0.344	0.030044

结果

		CI宽度的一半	下限（L）	上限（U）	
双侧		0.058885617	0.285114383	0.40288562	$(p-z_{\alpha/2}S_p,\ p+z_{\alpha/2}S_p)$
单侧	总体参数位于右尾	0.049418367	0.294581633	∞	$(p-z_{\alpha}S_p,\ +\infty)$
	总体参数位于左尾	0.049418367	-∞	0.39341837	$(-\infty,\ p+z_{\alpha}S_p)$

$S_p=\sqrt{\frac{p(1-p)}{n}}$

（2） 两总体率之差的CI

	置信度为95% α	n_1	p_1	n_2	p_2	p_C	$S_{p_1-p_2}$
请输入已知数据	0.05	80	0.75	70	0.571	0.6664667	0.077163

结果

		CI宽度的一半	下限（L）	上限（U）	
双侧		0.151237171	0.027762829	0.33023717	$(\lvert p_1-p_2\rvert-z_{\alpha/2}S_{p_1-p_2},\ \lvert p_1-p_2\rvert+z_{\alpha/2}S_{p_1-p_2})$
单侧	总体参数位于右尾	0.126922235	0.052077765	∞	$(\lvert p_1-p_2\rvert-z_{\alpha}S_{p_1-p_2},\ +\infty)$
	总体参数位于左尾	0.126922235	-∞	0.30592224	$(-\infty,\ \lvert p_1-p_2\rvert+z_{\alpha}S_{p_1-p_2})$

$S_{p_1-p_2}=\sqrt{p_C(1-p_C)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$　$p_C=\frac{n_1p_1+n_2p_2}{n_1+n_2}$

图 20-5　总体率的置信区间

实例与结果：例如，某卫生防疫站 2002 年对某乡镇 250 名小学生进行贫血的检测，结果发现有 86 名贫血者，检出率为 34.4%，求贫血检出率的 95%的置信区间。该数据属于“单个总体率的 CI”，所以在程序的第 5 行键入已知数据，得到总体率的 95%置信区间为

(0.2851, 0.4029)，即该小学生贫血检出率的95%置信区间为（28.51%,40.29%），反映精确度的置信区间宽度的一半为5.89%。如果研究需要获得单侧置信区间，该程序也已经给出。

温馨提示：该程序提供了获得单个总体率和两个总体率之差的置信区间估计程序，可根据需要选用。对于两个总体率的比较，还给出了一种简单计算的备选方法，多数情况下选择图20-5界面下方呈现的公式较好。

4. 总体方差的置信区间估计

当总体 $N(\mu,\sigma^2)$ 的参数 μ,σ^2 都未知时，方差 σ^2 的 $100(1-\alpha)\%$ 置信区间为：

$$\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2,(n-1)}},\frac{(n-1)S^2}{\chi^2_{(1-\alpha/2),(n-1)}}\right)$$

两个方差之比的置信区间为：

$$\left(\left(\frac{S_1^2}{S_2^2}\right)\frac{1}{F_{\alpha/2,\nu_1,\nu_2}},\ \left(\frac{S_1^2}{S_2^2}\right)\frac{1}{F_{1-\alpha/2,\nu_1,\nu_2}}\right)$$

其中 $\chi^2_{\alpha/2,(n-1)}$、$\chi^2_{(1-\alpha/2),(n-1)}$、$F_{\alpha/2,\nu_1,\nu_2}$、$F_{1-\alpha/2,\nu_1,\nu_2}$ 是在 χ^2 分布或 F 分布下的界值。

电脑实验

打开光盘文件“20-3 置信区间的估计.xlsx”的“4. 方差的置信区间”工作表（如图20-6所示），在“请输入已知数据”右侧键入数据，其下的“结果”会自动修正。

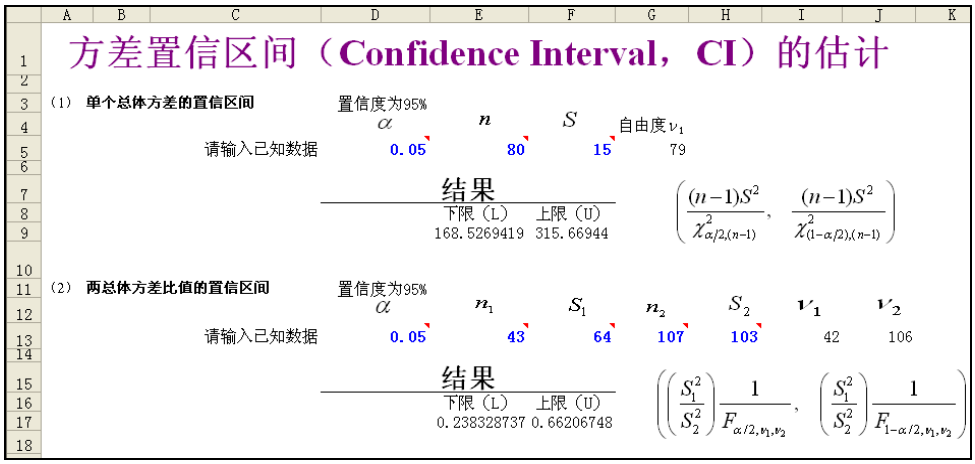


图20-6 总体方差的置信区间

实例与结果：例如，随机抽查某地区80名血吸虫病病人，测得血红蛋白均数为95.0g/L，

标准差为 15.00g/L，试估计总体方差的置信区间。该数据属于“单个总体方差的置信区间”，所以在程序的第 5 行键入已知数据，得到总体方差的 95%置信区间为（168.53, 315.67）。

温馨提示：该程序提供了获得单个总体方差和两个总体方差之比的置信区间估计程序，可根据需要选用。

20.3 假设检验的两型错误

因为处理的是随机样本，下统计学结论都会冒一定的错误风险。如果检验假设 H_0 正确，样本假设检验的推论结果拒绝正确的 H_0，则犯了 I 型错误（type I error），其错误的概率记为 α；如果 H_0 不正确，即总体间有实际差异，而假设检验没有拒绝不正确 H_0，则犯了 II 型错误（type II error），其错误的概率记为 β（参见表 20-1）。

表 20-1 假设检验的两型错误

	样本假设检验的推断结论	
真实情况	拒绝 H_0	不拒绝 H_0
H_0 正确	I 型错误且犯错误的概率为 α，即检验水准	推断正确 且正确结论的概率为 $1-\alpha$，即置信度
H_0 不正确	推断正确 且正确结论的概率为 $1-\beta$，即检验功效	II 型错误 且犯错误的概率为 β

如图 20-7 所示，电脑实验的零假设 H_0：$\mu=\mu_0=71.2$，即样本均数来自于 $\mu_0=71.2$ 的总体；备择假设 H_1：$\mu=\mu_1=74$，即样本均数来自于 $\mu_1=74$ 的总体。数据的第 A 列为假定的样本均数（如脉搏跳动的样本均数）；第 B 列为 H_0 的分布数据，H_0 的分布假定为 $N(\mu_0,\sigma_0^2/n)=N(71.2,8^2/82)$，在 B7 中键入公式“=NORMDIST(A7,$B$3,$B$5/SQRT($E$27),0)”，将此公式向下拖放至 B507，其中单元格 B3、B5 分别存放 μ_0、σ_0，单元格 E27（由滚动条控制）存放样本含量 n。第 C 列为 H_1 的分布数据，H_1 的分布假定为 $N(\mu_1,\sigma_1^2/n)=N(74,6^2/82)$，在 C7 中键入公式“=NORMDIST(A7,$E$25,$C$5/SQRT($E$27),0)”，将此公式向下拖放至 C507，其中单元格 E25（由滚动条控制）、C5 分别存放 μ_1、σ_1，单元格 E27 存放样本含量 n。采用 A7:A507 为横轴，B7:B507 为纵轴，绘制 H_0 分布的散点图；采用 A7:A507 为横轴，C7:C507 为纵轴，绘制 H_1 分布的散点图；采用 E23（由滚动条控制）制作临界值竖线。采用“开发工具→插入→表单控件”命令，在表单控件下拉窗口中，找到第二行第三个“滚动条（窗口控件）制作滚动条，略加修饰便可制成如图 20-7 所示的电脑实验。

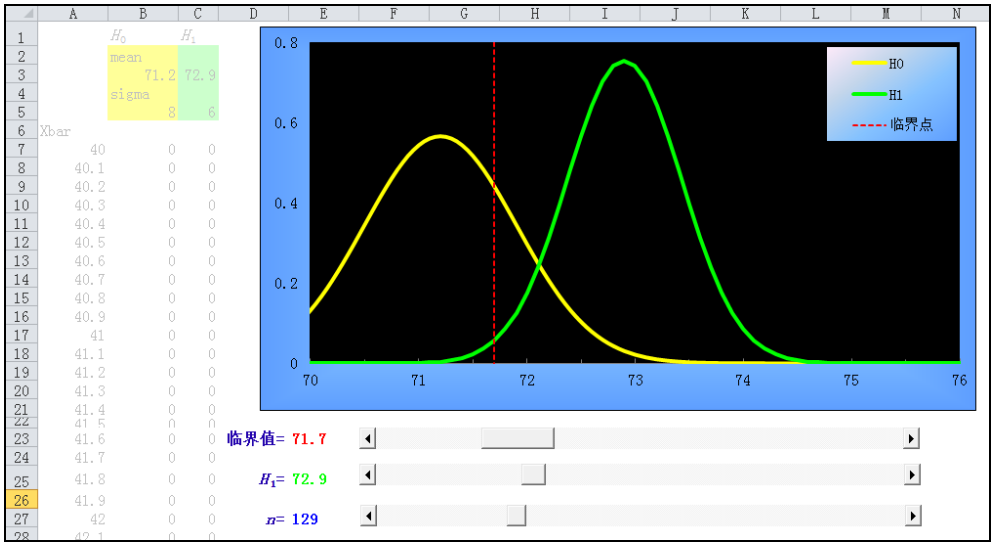


图 20-7 两型错误电脑实验

电脑实验

打开光盘文件“20-4 两型错误.xlsx”：

①左右滑动“临界值”滚动条，可以看到临界值左右移动，“临界值”一旦确定，那么I型错误便确定下来。在左侧 H_0 分布曲线下，临界值以右面积为I型错误概率 α，曲线下临界值左侧面积为置信度 $1-\alpha$；如果 H_1 对应的总体参数也已知，那么在右侧 H_1 分布曲线下，临界值以左面积为II型错误概率 β，曲线下临界值右侧面积为检验功效 $1-\beta$。假设检验是以 H_0 为前提条件，然后根据样本数据进行统计学推断，如果备择假设 H_1 分布未知，则无法知道II型错误概率 β 及检验功效 $1-\beta$。

②左右滑动 H_1 分布均数“$H1$”滚动条可见，两个总体差异 $|\mu_1-\mu_0|$ 越大，犯II型错误的机会越少，检验功效越大。

③左右滑动样本含量“n”滚动条可见，H_0 分布与 H_1 分布的曲线形状发生明显改变。样本含量越大两条曲线越高而瘦，α、β 两型错误概率越小；相反，样本含量越小两条曲线越矮而胖，α、β 两型错误概率越大。即增大样本含量可同时减少两型错误。

20.4 不同检验方法所导致结论的差异

统计学方法的选择很重要，不同设计、不同数据类型应选择不同的方法。

例 20-1：将 16 只大鼠配成 8 对，分别喂两种饲料一段时间后，测其肝中的维生素 A 的含量（IU/mg）如下：

对子号	1	2	3	4	5	6	7	8
饲料 1	3.55	3.80	2.00	3.75	3.00	3.45	3.95	3.05
饲料 2	2.45	3.25	2.40	2.70	2.80	2.50	3.20	3.10

问不同饲料喂养鼠肝中维生素 A 的含量之间的差别有无统计学意义？

该数据应该采用配对 t 检验进行分析，如果采用成组 t 检验结论将完全相反。配对 t 检验得到双侧 P 值为 0.0325，小于 0.05，说明差异有统计学意义；而成组 t 检验得到双侧 P 值为 0.0611，大于 0.05，说明差异无统计学意义（如图 20-8 所示）。

不同检验方法检验结果不同

编号	A	B	差值
1	3.55	2.45	1.1
2	3.8	3.25	0.55
3	2	2.4	-0.4
4	3.75	2.7	1.05
5	3	2.8	0.2
6	3.45	2.5	0.95
7	3.95	3.2	0.75
8	3.05	3.1	-0.05

配对 t 检验结果

	A组	B组
均数	3.3188	2.8000
标准差	0.6324	0.3454
方差	0.4000	0.1193
对子数	8	
Pearson相关系数	0.4922	
差值的标准误	0.1950	
df	7.0000	
t Stat	2.6601	
P(T<=t) 单尾	0.0162	
t 单尾临界	1.8946	
P(T<=t) 双尾	0.0325	
t 双尾临界	2.3646	

$$t=\frac{\bar{d}-\mu_{\bar{d}}}{S_{\bar{d}}}=\frac{\bar{d}}{S_d/\sqrt{n}},\quad \nu=\text{对子数}-1$$

成组 t 检验结果

	A组	B组
均数	3.3188	2.8000
标准差	0.6324	0.3454
方差	0.4000	0.1193
观察例数	8	8
合并方差	0.2596	
标准误	0.2548	
df	14.0000	
t Stat	2.0362	
P(T<=t) 单尾	0.0306	
t 单尾临界	1.7613	
P(T<=t) 双尾	0.0611	
t 双尾临界	2.1448	

$$t=\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{S_{\bar{X}-\bar{X}}}\qquad \nu=n_1+n_2-2$$

$$S_{\bar{X}-\bar{X}}=\sqrt{S_C^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}=\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}\left(\frac{n_1+n_2}{n_1n_2}\right)}$$

图 20-8　配对 t 检验与成组 t 检验

图 20-8 制作中所用的主要计算公式见图下方，具体参见第 8 章。

电脑实验

打开光盘文件“20-5 配对与成组 t 检验.xlsx”，修改单元格 C10 中的数据，如将“3.1”修改成“2”时，两种检验差异均有统计学意义（P 值分别为 0.0096、0.0283）；将“3.1”修改成“4”时，两种检验差异均无统计学意义（P 值分别为 0.1658、0.1905）。

温馨提示：单元格区域 B3:C102 是用户数据输入区。除做以上电脑实验外，该程序还可进行两组中每一组例数不超过 100 的配对 t 检验或成组 t 检验。

20.5 *t* 检验的两型错误

为了理解两型错误，下面通过正态分布随机函数产生随机数，然后采用 *t* 检验对随机数进行假设检验，并清点假设检验差异有无统计学意义的样本份数，换算成百分比，以此理解 I 型错误与 II 型错误的含义。

光盘文件“20-6 t 检验的两型错误.xlsx（如图 20-9 所示）共有“t 检验”、“随机数据”两个工作表，其中“随机数据”分成两大块，每块数据有 10 列、1000 行。左侧数据块的正态分布参数在“t 检验”工作表的单元格 B4、C4 内控制，每格的计算公式为“=NORMINV(RAND(), t 检验!$B$4, t 检验!$C$4)”；右侧数据块的正态分布参数在“t 检验”工作表的单元格 F4、G4 内控制，每格的计算公式为“=NORMINV(RAND(),t 检验!$F$4, t 检验!$G$4)”。在“t 检验”工作表中，单元格区域 C9:D1008 为“随机数据”左侧数据块的均数标准差，单元格区域 G9:H1008 为“随机数据”右侧数据块的均数标准差，单元格区域 J9:K1008 为左右两个数据块的 *t* 检验结果。单元格 L2 的计算公式为“=COUNTIF(K9:K1008,"<="&0.05)”，即清点了左右两个数据块的 *t* 检验 *P* 值小于等于 0.05（即差异有统计学意义）的样本份数，单元格 K2“=L2/1000”将有意义样本份数转换成了百分比，单元格 K3 为无统计学意义样本份数百分比。为了反映样本含量对假设检验统计学差异的影响，特在“随机数据”左、右侧数据块各 10 列数据中选前 5 列数据进行类似两份样本 *t* 检验，每份样本 *t* 检验结果存放在单元格区域 M9:N1008 中，单元格 L4、L5 分别是单元格区域 N9:N1008*P* 值小于等于 0.05、*P* 值大于 0.05 的样本份数，K4、K5 是样本含量为 5 的有、无统计学意义样本份数百分比。

P9:Q1008 为“随机数据”左、右侧数据块各 10 列数据的方差齐同性 *F* 检验，其有、无统计学意义样本份数以及百分比见 C5:D6。



电脑实验

打开光盘文件“20-6 t 检验的两型错误.xlsx”：

①在两个总体完全相同（均为 $N(150,20^2)$）的情况下，观察几个百分比的变化情况。在 H_0：$\mu_1=\mu_2$ 成立的情况下，拒绝 H_0，认为差异有统计学意义是犯了 I 型错误，也就是说，单元格 K2、K4 中的百分比值即代表了由样本获得的 I 型错误概率。按 F9 功能键，可以发现 K2、K4 中的百分比均在 $\alpha=0.05$ (5%)上下波动。

②任意修改一个总体均数，比如修改单元格 B4 中的“150”为 200，现在两个总体不同，分别为 $N(200,20^2)$、$N(150,20^2)$。在 H_0：$\mu_1=\mu_2$ 成立的情况下，不拒绝错误的 H_0，认为差异无统计学意义是犯了 II 型错误，也就是说，单元格 K3、K5 中的百分比值即代表

了由样本获得的 II 型错误概率；相应的单元格 K2、K4 是检验功效$1-\beta$。按 F9 功能键，可以发现 K2、K4 中的百分比相差较大，$n=10$ 对应的检验功效（单元格 K2）相对较大，围绕 99.87%（由文件 10-7 的第 2 个工作表计算）周围波动，$n=5$ 对应的检验功效（单元格 K4）相对较小，围绕 93.09%（由文件“20-7 t 检验与 t'检验.xlsx”的第 2 个工作表计算）周围波动。从而表明样本含量越大，检验功效越高。



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1																	
2			总体1				总体2			$n=10$	5.10%	51	份两样本差异有统计学意义				
3			均数 标准差				均数 标准差				94.90%	949	份两样本差异无统计学意义				
4			150 20				150 20			$n=5$	5.30%	53	份两样本差异有统计学意义				
5			5.00%	50	份两样本方差不齐						94.70%	947	份两样本差异无统计学意义				
6			95.00%	950	份两样本方差齐同												
7	样本									$n=10$			$n=5$			$n=10$	
8	编号	例数1	均数1	标准差1		例数2	均数2	标准差2		t值	P值		t值	P值		F值	P值
9	1	10	147.1966	21.8991		10	144.44	20.319		0.291939	0.773671		0.0418	0.9677		1.161531	0.827142
10	2	10	146.6328	27.1399		10	145.79	25.788		0.070955	0.944216		0.798	0.4479		1.107583	0.881516
11	3	10	145.382	20.22		10	128.84	22.05		1.748146	0.097474		0.4949	0.634		1.189238	0.800494
12	4	10	137.6677	21.03		10	155.9	23.118		1.845087	0.081543		1.3835	0.2039		1.208413	0.782549
13	5	10	144.3209	13.7197		10	138.76	13.108		0.926326	0.366531		1.3685	0.2084		1.095482	0.89417
14	6	10	157.1493	19.9688		10	148.29	13.474		1.162916	0.260049		1.0576	0.3211		2.196311	0.256791
15	7	10	165.4475	23.7738		10	143.5	25.893		1.974422	0.06388		2.0834	0.0707		1.186238	0.803338
16	8	10	142.2871	15.3871		10	154.77	25.802		1.314413	0.205211		0.3096	0.7648		2.811895	0.139515
17	9	10	149.5149	17.5666		10	151.17	11.241		0.251025	0.804638		0.2094	0.8394		2.441991	0.19967
18	10	10	149.121	18.3652		10	148.92	22.514		0.022086	0.982622		1.8795	0.097		1.502812	0.553633
19	11	10	158.3352	22.5954		10	150.12	18.589		0.888195	0.386151		0.9931	0.3497		1.477522	0.570171
20	12	10	150.4157	14.7601		10	155.69	18.239		0.710767	0.486337		0.2281	0.8253		1.526947	0.538336
21	13	10	150.3801	16.0645		10	142.39	18.679		1.025871	0.318542		1.2775	0.2372		1.351912	0.660582
22	14	10	153.0767	20.3746		10	147.69	20.242		0.592841	0.560663		0.3279	0.7514		1.013184	0.984754
23	15	10	143.5619	20.9052		10	154.44	19.224		1.211503	0.241368		1.6588	0.1357		1.182522	0.806874
24	16	10	145.5929	21.0791		10	149.3	21.174		0.392524	0.699281		0.3179	0.7587		1.009019	0.989548

t检验 / 随机数据 /

图 20-9　t 检验的两型错误

温馨提示：该程序也可查看标准差改变、总体方差不齐情况下的检验功效。

20.6　t 与 t' 检验

该实验与前面“t 检验的两型错误”的制作十分相似，不同的是：除对“随机数据”左、右侧数据块各 10 列数据进行两样本成组 t 检验以外，也对其进行 t' 检验。通过改变总体均数、总体标准差，比较不同方差情况下 t 检验与 t' 检验。

电脑实验

打开光盘文件“20-7 t 检验与 t'检验.xlsx”（如图 20-10 所示）：

①在两个总体完全相同（均为 $N(100,10^2)$）的情况下，两种检验的 I 型错误概率略有所不同。一般情况下，t 检验的 I 型错误概率（单元格 K2）高于 t' 检验（单元格 K4），t' 检验的实际检验水准 α 低于设定的 0.05。

②任意修改一个总体均数，比如修改单元格 B4 中的“100”为 90，使两个总体均数不同但方差相同，相应的单元格 K2、K4 是检验功效1－β。按 F9 功能键，可以发现t检验的检验功效（如 57%）高于t'检验（如 56.4%）。

③任意修改一个总体标准差，比如修改单元格 C4 中的“10”为 5，使两个总体均数、总体方差均不相同，相应的单元格 K2、K4 也不再是简单的检验功效1－β，即检验出的差异除了两个均数之差外，还有两个方差之差。按 F9 功能键，可以发现t检验使得单元格 K2 值高于单元格 K4。

不同方差情况下的t检验与t'检验

	总体1				总体2				t检验	5.20%	52	份两样本差异有统计学意义
	均数	标准差			均数	标准差				94.80%	948	份两样本差异无统计学意义
	100	10			100	10			t'检验	4.90%	49	份两样本差异有统计学意义
		5.50%		55	份两样本方差不齐					95.10%	951	份两样本差异无统计学意义
		94.50%		945	份两样本方差齐同							

样本编号	例数1	均数1	标准差1	例数2	均数2	标准差2	t值	P值	t'值	P值	F值	P值
1	10	100.443	11.09477	10	98.4304	9.92532	0.427535	0.67406	0.4275	0.6744	1.24953	0.7454
2	10	99.19356	8.017692	10	96.0236	8.0204	0.883923	0.388392	0.8839	0.3891	1.00068	0.9992
3	10	97.97371	8.817756	10	98.1849	5.93087	0.062849	0.950579	0.0628	0.9507	2.21045	0.253
4	10	97.67684	7.07341	10	103.813	10.4223	1.540399	0.140859	1.5404	0.1443	2.17104	0.2637
5	10	100.3435	6.638787	10	96.9927	7.63362	1.047397	0.308778	1.0474	0.3096	1.32216	0.6842
6	10	100.1298	11.08319	10	99.9088	5.24982	0.056984	0.955186	0.057	0.9555	4.45697	0.0364
7	10	95.64933	9.658605	10	100.923	5.58715	1.494529	0.152362	1.4945	0.1572	2.98847	0.1185
8	10	107.6081	7.247047	10	103.483	10.3593	1.031846	0.31581	1.0318	0.3175	2.04333	0.302
9	10	103.1707	8.650545	10	98.8483	10.6224	0.997783	0.33161	0.9978	0.3324	1.50784	0.5504
10	10	100.9452	11.34986	10	97.4223	11.5662	0.687472	0.500548	0.6875	0.5011	1.03849	0.9561
11	10	103.4628	6.992861	10	104.5	12.4033	0.230355	0.820414	0.2304	0.8211	3.14603	0.1029
12	10	103.227	6.149593	10	102.306	9.35714	0.260104	0.797736	0.2601	0.7983	2.31523	0.227
13	10	93.61957	5.223696	10	100.461	13.1856	1.525357	0.144549	1.5254	0.1554	6.37149	0.011
14	10	103.9541	9.36086	10	105.717	9.6978	0.413677	0.683998	0.4137	0.6843	1.07328	0.9178
15	10	107.5657	11.99765	10	102.294	10.8996	1.028462	0.317355	1.0285	0.3181	1.21164	0.7796
16	10	100.4821	9.147879	10	98.7564	16.395	0.290676	0.774622	0.2907	0.7756	3.21207	0.0971

图 20-10 t检验与t'检验

温馨提示：除样本需来自独立、随机的正态分布以外，t检验还需满足方差齐同的条件，如果方差不齐，很难得出差异是两个总体均数之间的差异。所以在方差不齐的情况下，需要采用校正了自由度的t'检验。

20.7 一般 t 检验做方差分析两两比较导致的 I 型错误

光盘文件“20-8 一般 t 检验做方差分析两两比较导致的 I 型错误.xlsx”（如图 20-11 所示）有“方差分析”、“随机数据”、“两两比较”三个工作表，其中“随机数据”为 10 列、100 行，每格的公式为“=NORMINV(RAND(), 方差分析!$M$3, 方差分析!$N$3)”，“方差分析”单元格 M3、N3 可控制“随机数据”的正态分布参数。“方差分析”工作表中单元格区域 C4:C103、D4:D103 分别为“随机数据”每行 10 个数字的均数、标准差，前 100 份样

本的前 10 份样本以及全部样本的误差条图见单元格区域 F10:R15，对应的方差分析表见单元格区域 F5:R9。100 份样本的两两比较 t 检验 P 值存放在“两两比较”工作表的单元格区域 B4:CV102 中。在“方差分析”工作表的单元格区域 F17:K22 与单元格区域 M17:R22 中，分别显示了前 10 份（共计比较 $C_{10}^2=45$）随机样本与全部 100 份（共计比较 $C_{100}^2=4950$）样本的两两比较 t 检验情况。

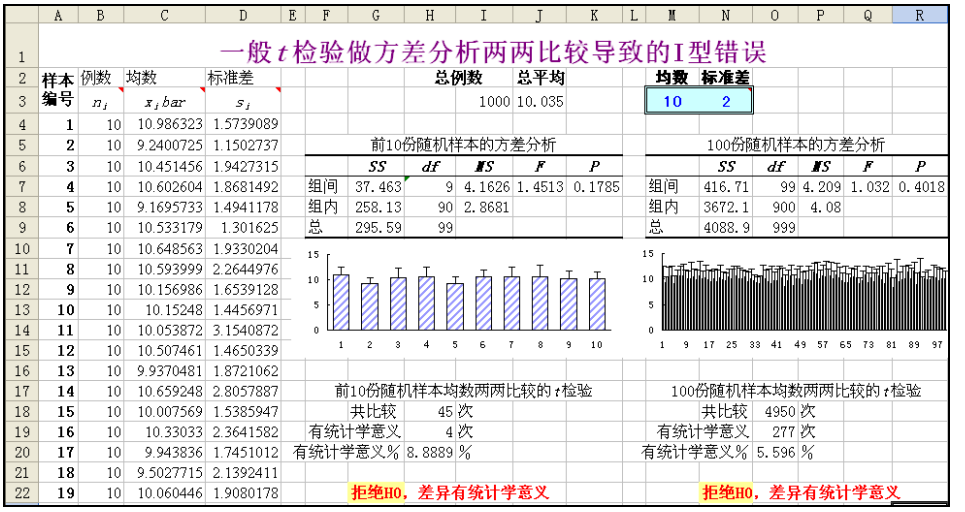


图 20-11　一般 t 检验做方差分析两两比较导致的 I 型错误

电脑实验

打开光盘文件“20-8 一般 t 检验做方差分析两两比较导致的 I 型错误.xlsx”，按 F9 功能键，观察每次新生成的来自同一个正态总体的随机数方差分析与 t 检验结果。

因为是随机样本，存在抽样误差，会出现第一类错误。按 α =0.05 检验水准，方差分析有统计学意义的频率在 5%左右，无论是前 10 份样本还是全部 100 份样本都一样。

而对每份样本进行两两比较 t 检验的结果不同。差异有统计学意义的百分比远远超过 5%，说明多组均数比较用 t 检验加大了 I 型错误概率，而且两两比较的次数（即样本处理分组组数）越多，I 型错误越严重。10 份样本（共计 45 次）两两比较有约 50%以上的 I 型错误概率，100 份样本（共计 4950 次）两两比较接近 100%的 I 型错误概率。

20.8　样本含量对假设检验的影响

在比较如图 20-12 所示表格中的三个率（24.68%、15.56%和 16.40%）时，分析结果表

明：第一个率和第二个率之间差别（9.13%）无统计学意义（P=0.0865）；尽管第一个率和第三个率之间差值只有 8.29%，但检验结果反而有统计学意义（P=0.0006）。这个结果使很多人不理解，怀疑是否计算机软件出错。究其原因是样本含量引起，假设检验是根据抽样误差大小进行推断的，样本含量越大，抽样误差越小。如定量数据的抽样误差（均数标准误）由公式 $S_{\bar{X}} = S/\sqrt{n}$ 计算，定性数据的抽样误差（率的标准误）由公式 $S_p = \sqrt{p(1-p)/n}$ 计算。检验统计量如 t、z 等计算公式一般为：

$$\text{检验统计量}=\frac{\text{样本差值}}{\text{抽样误差}}$$

形式。产生第一个率和第二个率的例数之和只有 438；而产生第一个率和第三个率的例数之和有 1009，是前者的两倍多，所以导致了两率之差大没有统计学意义，两率之差小反而有统计学意义的现象。

	A	B	C	D	E	F	G	H
1		**阳性**	**阴性**	**合计**	**阳性率**			
2	率1	97	296	393	0.2468			
3	率2	7	38	45	0.1556			
4	率3	101	515	616	0.1640			
5								
6		两率之差	合并率	合计例数	标准误	z值	P值	卡方值
7	率1与率2	0.0913	0.237442922	438	0.066965377	1.36285	0.0865	1.857361
8	率1与率3	0.0829	0.196233895	1009	0.025639574	3.231657	0.0006	10.4436
9	率2与率3	-0.0084	0.163388805	661	0.057092242	0.147226	0.4415	0.021676

图 20-12 率比较实例

光盘文件“20-9 样本含量对假设检验的影响.xlsx”（如图 20-13 所示）通过简单四格表卡方检验，比较不同样本含量倍数对假设检验结果的影响。制作方法是：令四格表第 1 个格子频数可变，通过滚动条控制在 1～100 之间变化，四格表其他 3 个格子的频数固定为 30。右侧四格表格子频数是左侧四格表格子频数的倍数，倍数大小通过滚动条控制在 1～100 之间。对于每一四格表的假设检验结果见相应表格的下方。

	A	B	C	D	E	F	G	H	I	J	K	L
1	样本含量对样本率假设检验结果的影响											
2												
3			**假想四格表数据**					**每一格子数字是假想四格表的**	100	**倍**		
4		分组	例数		合计	阳性率(%)		分组	例数		合计	阳性率(%)
5			阳性	阴性					阳性	阴性		
6		实验组	34	30	64	53.13		实验组	3400	3000	6400	53.13
7		对照组	30	30	60	50.00		对照组	3000	3000	6000	50.00
8		合计	64	60	124	51.61		合计	6400	6000	12400	51.61
9												
10			χ^2=	0.121	P值	0.7279			χ^2=	12.11	P值	0.0005
11												
12												
13				实验组阳性数								
14				倍数								

图 20-13 率比较电脑实验

电脑实验

①固定倍数滚动条为 100，左右滑动频数（实验组阳性数）滚动条，观察假设检验结果的变化。当滑到实验组阳性数为 34 时，两个四格表两组阳性率分别为 53.13%、50.00%。左侧四格表卡方检验 P=0.7279，100 倍时右侧四格表卡方检验 P=0.0005。

②固定频数滚动条为 34，左右滑动倍数滚动条，观察假设检验结果的变化。当滑到倍数为 30 时，左侧四格表卡方检验 P=0.7279，30 倍时右侧四格表卡方检验 P=0.0567，按 α =0.05 检验水准，差异均无统计学意义。

结果分析：从上述可见，样本含量越大，假设检验越有可能获得有统计学差异的结论。

20.9 卡方检验公式校正的意义

在四格表资料中，$N\geqslant 40$，但是有一个格子的理论频数 $1\leqslant T<5$ 时，利用四格表 χ^2 检验公式计算出来的 χ^2 统计量偏大，应加以校正。英国统计学家 Yates F 提出将实际频数与理论频数之差的绝对值减去 0.5 作连续性校正（correction for continuity），故连续性校正公式又称 Yates 校正（Yates's correction）公式。χ^2 检验连续性校正公式为：

$$\chi^2=\sum\frac{(|A-T|-0.5)^2}{T}$$

从相同二项分布 $B(n,p)$ 中，分别产生 a_0、b_0 两个格子数据，两组合计例数均为 n（如表 20-1 所示）。因为 A、B 两组例数相同，所以其理论频数 $T_{b_0}=T_{a_0}$ 、$T_{b_1}=T_{a_1}$，如果知道 T_{a_0} 、T_{a_1} 便可知道其他两个。四个格子理论频数只要有一个小于 5，则需要校正。检验水准 α =0.05。

表 20-1

分　组	阳　性	阴　性	合　计
A	a_0	$a_1=n-a_0$	n
B	b_0	$b_1=n-b_0$	n
合计	a_0+b_0	a_1+b_1	$N=2n$

根据以上原则制作的程序如图 20-14 所示。其中 B、C 两列各 1000 个数据由二项分布 $B(30,0.15)$ 控制（见单元格 O6、O7），如 B14 为“=CRITBINOM($O$6,$O$7,RAND())”。第 D、E 列由 B、C 列获得，如 D14 为“=$O$6-B14”，E14 为“=$O$6-C14”。第 F、G 列的 Ta0、Ta1 为理论频数 T_{a_0} 、T_{a_1}，第 H 列“应校正”就是根据第 F、G 列而得，1 表示

需要校正，0 为无须校正。单元格区域 J14:J1013、N14:P1013 分别为一般卡方检验、校正卡方检验结果，其中 1 表示在 α =0.05 水准有统计学意义，0 表示无统计学意义。为了清点校正与否的统计学意义频数，特别在第 R 列根据 H14:H1013、L14:L1013、P14:P1013 三列产生分类类别，单元格 D6、E6、D7、E7、J6、K6、J7、K7 分别对应于 1、2、3、4、5、6、7、8。

电脑实验

打开光盘文件“20-10 卡方检验公式校正的意义.xlsx”，每次按 F9 功能键，将重新生成 1000 个随机四格表。是否需要校正数量及配对四格表检验结果都会随机改变。

结果分析：

①每次一般卡方检验的差异有统计学意义的概率总是超过校正卡方检验。表明如果需要校正而没有校正的话，可能增大 I 型错误的概率。

②在二项分布为 $B(30,0.15)$ 情况下，需要校正四格表数约为 60%，不需要校正四格表数约为 40%；无论是在需要校正的四格表中，还是在不需要校正的四格表中，两种卡方检验方法无统计学意义，配对四格表假设检验差异均有统计学意义，表明这两种方法得出的检验结果是不一致的。因此，统计分析前应事先考虑检验方法的应用条件，否则会得出不正确结论。

一般卡方检验和校正卡方检验的结果比较

40%	399个四格表不需要校正			60%	601个四格表需要校正		
	卡方检验	校正卡方检验	合计		卡方检验	校正卡方检验	合计
		有意义 无意义				有意义 无意义	
	有意义	11 9	20		有意义	9 24	33
	无意义	0 379	379		无意义	0 568	568
	合 计	11 388	399		合 计	9 592	601
配对卡方值	9	P值 =	0.0027	配对卡方值	24	P值=	<0.0001
	一般卡方检验	校正卡方检验			一般卡方检验	校正卡方检验	
	有意义率	5.01%	2.76%		有意义率	5.49%	1.50%

$B(n,\pi)$：n = 30，π = 0.15

编号	a0	b0	a1	b1	Ta0	Ta1	应校正	Pearson 卡方	P 值	统计意义	校正 卡方	P 值	统计意义	类别
1	5	6	25	24	5.5	24.5	0	0.11132	0.738649	0	0	1	0	4
2	2	6	28	24	4	26	1	2.30769	0.128735	0	1.298077	0.25456	0	8
3	5	2	25	28	3.5	26.5	1	1.45553	0.227644	0	0.6469	0.42122	0	8
4	6	3	24	27	4.5	25.5	1	1.17647	0.278076	0	0.522876	0.46962	0	8
5	4	4	26	26	4	26	1	0	1	0	0.144231	0.70411	0	8
6	5	7	25	23	6	24	0	0.41667	0.518605	0	0.104167	0.74689	0	4
7	5	5	25	25	5	25	0	0	1	0	0.12	0.72903	0	4
8	4	5	26	25	4.5	25.5	1	0.13072	0.717688	0	0	1	0	8
9	3	6	27	24	4.5	25.5	1	1.17647	0.278076	0	0.522876	0.46962	0	8
10	5	5	25	25	5	25	0	0	1	0	0.12	0.72903	0	4

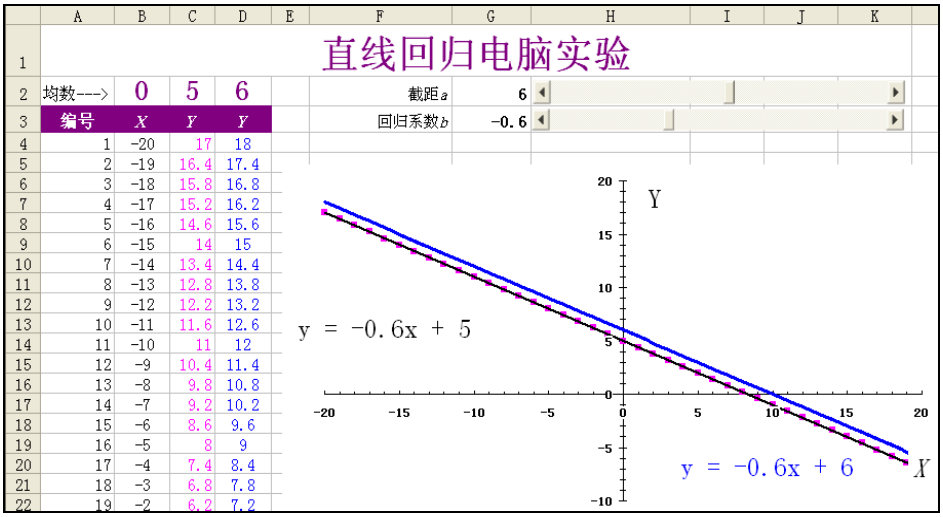
图 20-14 一般卡方检验与校正卡方检验结果的比较

温馨提示：在单元格 O6、O7 中，分别对应二项分布的参数 n、P，用户可以根据需要对此进行修改，如将 O6 中的“30”修改为“26”、“50”等，然后观察结果的变化。

20.10 回归参数意义理解

在第 10 章已经介绍了直线回归方程为 $\hat{Y}=a+bX$，式中 a 为回归直线在 Y 轴上的截距，a=0、a>0 和 a<0 分别表示回归直线与纵轴的交点在原点、原点之上和原点之下。b 为回归系数，即直线的斜率，表示变量 X 每增加（或减少）一个单位，Y 平均改变 b 个单位。

图 20-15（光盘文件“20-11 回归参数意义理解.xlsx”）的第 A 列为编号，B 列中 B4:B44 为假定的 X 变量，从–20 到+20，步长为 1。C 列中 C4:C44 为假定的（固定直线）Y 变量，从 17 到–7，步长为–0.6。D 列是动态直线的 Y 变量，其参数由 G2、G3 控制，G2、G3 又由 H2、H3（控件工具箱制作的滚动条）控制。其中 G2 的计算公式为“=(H2-20)”，当 H2 在 10 至 40 之间变化时，G2 将在–10 至 20 之间变化；其中 G3 的计算公式为“=(H3-20)/10”，当 H3 在 0 至 40 之间变化时，G2 将在–2 至 2 之间变化，步长为 0.1。数据右侧的图形为左侧数据对应的散点图，为了显示相应直线的回归方程，采用了“在相应直线散点上单击鼠标右键→设置趋势线格式→趋势线选项→选择显示公式”命令。



编号	X	Y	Y
1	-20	17	18
2	-19	16.4	17.4
3	-18	15.8	16.8
4	-17	15.2	16.2
5	-16	14.6	15.6
6	-15	14	15
7	-14	13.4	14.4
8	-13	12.8	13.8
9	-12	12.2	13.2
10	-11	11.6	12.6
11	-10	11	12
12	-9	10.4	11.4
13	-8	9.8	10.8
14	-7	9.2	10.2
15	-6	8.6	9.6
16	-5	8	9
17	-4	7.4	8.4
18	-3	6.8	7.8
19	-2	6.2	7.2

图 20-15　直线回归参数的意义

电脑实验

打开光盘文件“20-11 回归参数意义理解.xlsx”：

①当左右滑动“截距”滚动条时，直线将左右平行移动，在 Y 轴上的交点至原点（0,0）的距离即为截距值 a；动态直线对应的回归方程截距项发生改变。

②当左右滑动“回归系数”滚动条时，直线将围绕（ 0,a ）旋转，斜率发生改变；动态直线对应的回归方程回归系数项发生改变。

③将 X 值改变，即制造出“异常值”（ outlier ）（ 如将第一个 X 值“-20”修改为 20 ），可见回归直线拟合趋势线发生较大改变。

20.11　寿命表电脑实验

第 13 章已详细阐述了寿命表的制作。这里介绍内容的主要特点是：为了将实验结果与原结果形成对照，特地将第 J 列“期望寿命”选择性粘贴“数值”到第 L 列。这样第 J 列改变，第 L 列不变，由此形成对照。

电脑实验

打开光盘文件“20-12　寿命表电脑实验.xlsx”：

①左右滑动“0 岁组”死亡数（ 单元格 C3 ）滚动条，观察第 J 列期望寿命，可以发现：0 岁组期望寿命可能低于 1 岁组期望寿命，如“0 岁组”死亡数为 1000 时，0 岁组期望寿命为 67.59，小于 1 岁组期望寿命。“0 岁组”死亡数改变，其相应期望寿命发生改变，但其他各组期望寿命不发生改变。

②左右滑动“70 岁组”死亡数（ 单元格 C18 ）滚动条，观察第 J 列期望寿命，可以发现：70 岁及以下各组的期望寿命发生改变，但大于 70 岁组期望寿命不发生改变。此外图 20-16 中右侧图形也会发生相应改变。

现时简略寿命表电脑实验

年龄组（岁）	平均人口数（$_iP_x$）	死亡数（$_iD_x$）	年龄组死亡率（$_in_x$）	死亡概率（$_iq_x$）	尚存人数（l_x）	死亡人数（$_id_x$）	生存人年数（$_iL_x$）	生存总人年数（T_x）	期望寿命（e_x）	(ex)
0~	30005	429		0.014298	100000	1430	98785	6891768	68.92	68.92
1~	86920	105	0.001208	0.004820	98570	475	393331	6792984	68.92	68.92
5~	102502	81	0.000790	0.003943	98095	387	489508	6399653	65.24	65.24
10~	151494	113	0.000746	0.003723	97708	364	487632	5910145	60.49	60.49
15~	182932	157	0.000858	0.004282	97345	417	485681	5422513	55.70	55.70
20~	203107	215	0.001059	0.005279	96928	512	483359	4936832	50.93	50.93
25~	190289	221	0.001161	0.005790	96416	558	480685	4453472	46.19	46.19
30~	147076	181	0.001231	0.006134	95858	588	477819	3972788	41.44	41.44
35~	99665	160	0.001605	0.007995	95270	762	474445	3494969	36.68	36.68
40~	90891	234	0.002575	0.012790	94508	1209	469519	3020524	31.96	31.96
45~	105382	417	0.003957	0.019591	93299	1828	461927	2551006	27.34	27.34
50~	86789	602	0.006936	0.034091	91471	3118	449561	2089079	22.84	22.84
55~	69368	919	0.013248	0.064117	88353	5665	427603	1639518	18.56	18.56
60~	51207	1328	0.025934	0.121775	82688	10069	388268	1211914	14.66	14.66
65~	39112	1691	0.043235	0.195088	72619	14167	327677	823647	11.34	11.34
70~	20509	1561	0.076113	0.319726	58452	18689	245538	495970	8.49	8.49
75~	9301	1126	0.121062	0.464675	39763	18477	152624	250433	6.30	6.30
80~	3463	631	0.182212	0.625930	21286	13324	73122	97809	4.59	4.59
85~	834	269	0.322542	1.000000	7963	7963	24687	24687	3.10	3.10

0岁组

70岁组

问题1: 现时简略寿命表需要何种资料，如何获得期望寿命？

问题2: 0岁组预期寿命有可能低于1岁组期望寿命吗？

问题3: 0岁组期望寿命受其它年龄组死亡率的影响吗？

问题4: 随着年龄的增加，尚存人数、死亡人数、死亡概率、期望寿命各呈什么变化？

图 20-16　寿命表电脑实验

附录 A

Excel 函数

1. 数学和三角函数

（1）数学函数（math functions）

编号	函数	举例	说明（最末括号内为计算结果）
1	ABS(number)	=ABS(-2)	求-2 的绝对值（2）
2	AGGREGATE(function_num, options, ref1, [ref2], …)	=AGGREGATE(4,4,A1:A11)	返回列表或数据库中的聚合(A1:A11 区域中忽略空值后的最大值)
3	CEILING(number, significance)	=CEILING(2.005,2) 或=CEILING(-2.005,-2)	将 2.005 的绝对值向上取舍到最接近 2 绝对值倍数的整数（注意两个数符号相同）（4 或-4）
4	CEILING.PRECISE(number, significance)	=CEILING.PRECISE (2.005,2)或 =CEILING.PRECISE (-2.005,-2)	返回一个数字，该数字向上舍入为最接近的整数或最接近的有效位的倍数。无论该数字的符号如何，该数字都向上舍入。但是，如果该数字或有效位为零，则将返回零。（4 或-2）
5	COMBIN(number, number_chosen)	=COMBIN(8,2)	$C=\binom{n}{x}=\frac{n!}{x!(n-x)!}=\frac{8!}{2!(8-2)!}$ 从 8 个对象中提取 2 个对象的可能组合数（28）
6	EVEN(number)	=EVEN(1.41)或 =EVEN(-1.41)	将 1.41 沿绝对值增大的方向向上取舍到最接近的偶数（2 或-2）
7	EXP(number)	=EXP(1)	以自然对数 e 为底数的 1 次幂（2.718282）
8	FACT(number)	=FACT(5)	5 的阶乘，即 5!=5×4×3×2×1 的值。如果 Number 不是整数，则截尾取整（120）
9	FACTDOUBLE(number)	=FACTDOUBLE(5)	计算 5 双倍阶乘的数值。如果 Number 不是整数，则截尾取整(15)
10	FLOOR(number, significance)	=FLOOR(2.005,2)或 = FLOOR(-2.005, -2)	将 2.005 的绝对值向下取舍到最接近 2 绝对值倍数的整数（注意两个数符号相同）（2 或-2）
11	FLOOR.PRECISE(number, significance)	= FLOOR.PRECISE (2.005, 2) 或 = FLOOR.PRECISE (-2.005, -2)	返回一个数字，该数字向下舍入为最接近的整数或最接近的 significance 的倍数。无论该数字的符号如何，该数字都向下舍入。但是，如果该数字或有效位为零，则将返回零。
12	GCD(number1,[number2, ...])	=GCD(24, 36，12)	求 24、36 和 12 的最大公约数（仅在数据分析工具安装后才可用）（12）
13	INT(number)	=INT(8.9)或=INT(-8.9)	将正数 8.9 向下取舍到最接近的整数（8） 将负数-8.9 向下取舍到最接近的整数（-9）
14	LCM(number1,[number2, ...])	=LCM(5, 2，3)	求 5、2 和 3 的最小公倍数（仅在数据分析工具安装后才可用）（30）

续表

编号	函数	举例	说明（最末括号内为计算结果）
15	LN(number)	=LN(86)	求 86 的自然对数（4.454347）
16	LOG(number,[base])	=LOG(8, 2)	求以 2 为底 8 的对数（3）
17	LOG10(number)	=LOG10(86)	求以 10 为底 86 的对数（1.934498451）
18	MDETERM(array)	=MDETERM(A1:D4) =MDETERM({1,-1,1,-3;1,2,1,3;2,1,-1,-3;3,-3,-3,-1})	数组常量的矩阵行列式值（-126） 参见本书 2.9.3 节第（4）点
19	MINVERSE(array)	=MINVERSE(A1:C3) =MINVERSE({2,1,1;1,0,2;3,1,2})	求矩阵的逆（$\begin{pmatrix}-2 & -1 & 2\\ 4 & 1 & -3\\ 1 & 1 & -1\end{pmatrix}$） 参见 2.9.3 节第（3）点
20	MMULT(array1,array2)	=MMULT(A1:C2,E1:F3)=MMULT({1,0,3;2,1,0},{4,1;-1,1;2,0})	求矩阵的乘积$\begin{pmatrix}10 & 1\\ 7 & 3\end{pmatrix}$ 参见 2.9.3 节第（2）点
21	MOD(number,divisor)	=MOD(3,2)	3/2 的余数（1）
22	MROUND(number,multiple)	=MROUND(10, 3)	将 10 四舍五入到最接近基数 3 的倍数（9） 仅在数据分析工具安装后才可用
23	MULTINOMIAL(number1, number2, ...)	=MULTINOMIAL(2, 3, 4)	2、3、4 之和的阶乘与各自阶乘乘积的比值，即$\frac{(a+b+c)!}{a!b!c!}=\frac{(2+3+4)!}{2!3!4!}$（1260）
24	ODD(number)	=ODD(1.5)	将 1.5 向上取舍到最接近的奇数（3）
25	PI()	=PI()	求圆周率 π 的值(3.14159265358979)
26	POWER(number,power)	=POWER(98.6,3.2)	98.6 的 3.2 次方（2401077）
27	PRODUCT(number1,number2, ...)	=PRODUCT(5, 10, 100)	所给数据的连乘积，即 5×10×100（5000）
28	QUOTIENT(numerator, denominator)	=QUOTIENT(5, 2)	5/2 的商的整数部分（2）
29	RAND()	=RAND()	获得 0～1 之间的一个随机数（变量）
30	RANDBETWEEN(bottom,top)	=RANDBETWEEN(1, 100)	获得 1～100 之间的一个随机数（变量）
31	ROMAN(number,form)	=ROMAN(2013,0)	2013 的古典罗马数字样式（MMXIII）
32	ROUND(number, num_digits)	=ROUND(202.149, 1) =ROUND(202.149, -1)	将 202.149 四舍五入，小数位数为 1（202.1）； 小数位数为-1（200）
33	ROUNDDOWN(number, num_digits)	=ROUNDDOWN(3.2, 0)	将 3.2 向下取舍，小数位数为 0（3）
34	ROUNDUP(number, num_digits)	=ROUNDUP(3.2,0)	将 3.2 向上取舍，小数位数为 0（4）
35	SERIESSUM(x,n,m, coefficients)	=SERIESSUM(A2,0,2,A3:A6)	返回幂级数之和，SERIESSUM(x,n,m,a) $=a_1^n+a_2^{n+m}+a_3^{n+2m}+\cdots+a_j^{n+(j-1)m}$

续表

编号	函数	举例	说明（最末括号内为计算结果）
36	SIGN(number)	=SIGN(-0.00001)	负数的符号（-1） 正数结果为 1，负数结果为-1
37	SQRT(number)	=SQRT(16)	16 的平方根（4）
38	SQRTPI(number)	=SQRTPI(3)	求 $\sqrt{n\pi}=\sqrt{3\pi}$ =SQRT(3*PI())的值（3.06998）
39	SUBTOTAL(function_num, ref1,ref2,...)	=SUBTOTAL(9, A2:A5)	返回单元格区域 A2:A5 的分类汇总（fun_num 可取 1～11，分别表示 AVERAGE、COUNT、COUNTA、MAX、MIN、PRODUCT、STDEV、STDEVP、SUM、VAR、VARP）
40	SUM(number1,number2, ...)	=SUM(3,2)	将 3 和 2 相加（5）；也可以是单元格引用
41	SUMIF(range,criteria, sum_range)	=SUMIF(A2:A5,">160",B2:B5)	条件求和，参见 2.9.1 节第 9 点
42	SUMIFS(sum_range,criteria_range1,criteria1, [criteria_range2, criteria2], ...)	=SUMIFS(A1:A20, B1:B20, ">0", C1:C20, "<10")	对区域 A1:A20 中符合以下条件的单元格的数值求和：B1:B20 中的相应数值大于零 (0) 且 C1:C20 中的相应数值小于 10
43	SUMPRODUCT(array1,array2,array3, ...)	=SUMPRODUCT(A2:B4, C2: D4)	数组间对应元素乘积之和。如 3×2+4×7+8×6+6×7+1×5+9×3=156
44	SUMSQ(number1,number2, ...)	=SUMSQ(3,4)	3 与 4 的平方和（25＝3^2 + 4^2）；也可以是单元格引用，公式：$\sum X^2$
45	SUMX2MY2(array_x,array_y)	=SUMX2MY2(A2:A8,B2:B8)	两个数组对应元素的平方差之和； 公式为：$\sum\left(X^2-Y^2\right)$
46	SUMX2PY2(array_x,array_y)	=SUMX2PY2(A2:A8,B2:B8)	两个数组对应元素的平方和之和； 公式为：$\sum(X^2+Y^2)$
47	SUMXMY2(array_x,array_y)	=SUMXMY2(A2:A8,B2:B8)	两个数组中对应数值之差的平方和； 公式为：$\sum(X-Y)^2$
48	TRUNC(number,num_digits)	=TRUNC(8.9)	将 8.9 截断取整（8），参见 2.9.1 节

（2）三角函数（trigonometric functions）

编号	函　数	举　例	说明（最末括号内为计算结果）
1	ACOS(number)	=ACOS(-0.5)	以弧度表示-0.5（取值范围为-1～1）的反余弦值（2.094395）
2	ACOSH(number)	=ACOSH(10)	10（>1 的实数）的反双曲余弦值（2.993223）

续表

编号	函　数	举　例	说明（最末括号内为计算结果）
3	ASIN(number)	=ASIN(-0.5)	以弧度表示-0.5（取值范围为-1～1）的反正弦值（-0.5236）
4	ASINH(number)	=ASINH(-2.5)	-2.5 的反双曲正弦值（-1.64723）
5	ATAN (number)	=ATAN(1)	以弧度表示 1 的反正切值（0.785398）
6	ATAN2(x_num,y_num)	=ATAN2(1,1)	以弧度表示点（1,1）到原点的直线与 x 轴夹角的反正切值（0.785398）
7	ATANH(number)	=ATANH(-0.1)	-0.1 的反双曲正切值（-0.10034）
8	COS(number)	=COS(1.047)	1.047 弧度的余弦值（0.500171）
9	COSH(number)	=COSH(4)	4 的双曲余弦值（27.30823）
10	DEGREES(angle)	=DEGREES(PI())	圆周率 π 弧度转换为度数（180）
11	RADIANS(angle)	=RADIANS(270)	将角度 270 度转换为弧度（4.712389）
12	SIN(number)	=SIN(PI()/2)	π/2 弧度的正弦值（1）
13	SINH(number)	=SINH(1)	1 的双曲正弦值（1.175201194）
14	TAN(number)	=TAN(0.785)	0.785 弧度的正切值（0.99920）
15	TANH(number)	=TANH(-2)	-2 的双曲正切值（-0.96403）

2. 统计函数（statistical functions）

编号	函　数	举　例	说明（最末括号内为计算结果）
1	AVEDEV(number1,number2,　...)	=AVEDEV(A2:A8)	一组数据与其均值之差的绝对值的平均值 公式：$\sum\lvert X-\bar{X}\rvert/n$
2	AVERAGE(number1,number2, ...)	=AVERAGE(1,2,3,4,5)	获得一组数据（或单元格引用）的均数（3） $\sum X/n=\frac{1+2+3+4+5}{5}$
3	AVERAGEA(value1,value2,　...)	=AVERAGEA(A1:A6)	获得一组数据的均数；计算分母时，除对数值计数以外，还包括对文本、逻辑值等的计数。如果 A1:A5 为数字 1～5，A6 为“X”则计算结果为 $2.5=(1+2+3+4+5)/6$
4	AVERAGEIF(range,criteria,[average_range])	=AVERAGEIF(A2:A5,"<95000")	求在 A2:A5 区域内所有小于 95,000 的数值的算术平均值
5	AVERAGEIFS(average_range,criteria_range1,criteria1,[criteria_range2, criteria2], ...)	=AVERAGEIFS(C2:C5, C2:C5, ">95")	计算所有学生第二次测验成绩超过 95 分的分数的平均值。因为均未超过 95 分，所以返回 #DIV0!
6	BETA.DIST(x,alpha,beta,A, B)	=BETA.DIST(A2,A3, A4,A5,A6)	获得 Beta 分布的累积分布函数值；A、B 分别为数值 x 所属区间的可选下、上界

续表

编号	函　数	举　例	说明（最末括号内为计算结果）
7	BETA.INV(probability, alpha,beta,A,B)	=BETA.INV(A2,A3, A4,A5,A6)	Beta 分布的累积分布函数的反函数
8	BINOM.DIST(number_s, trials,probability_s, cumulative)	=BINOM.DIST(6,10, 0.3, 0)	总体率为 0.3 时 10 次试验中成功 6 次的概率（参见 17.2 节） Cumulative=0 或 FALSE 求概率质量函数 Cumulative=1 或 TRUE 求累积分布函数
9	BINOM.INV(trials,probability_s,alpha)	=BINOM.INV(10,0.7,0.05)	对于二项式分布 B（10,0.7），从左侧累积到曲线下面积为 0.05 时的最小临界值为 5。
10	CHISQ.DIST(x,deg_freedom,cumulative)	=CHISQ.DIST(18.307, 10,FALSE)	卡方值 18.307 在自由度为 10 的卡方分布中的概率密度函数值（0.0155）
11	CHISQ.DIST.RT(x,deg_freedom)	=CHISQ.DIST.RT(18.307, 10)	卡方值 18.307 在自由度为 10 的卡方分布中的 *P* 值（0.05）；参见 19.4.2 节
12	CHISQ.INV(probability, degrees_freedom)	=CHISQ.INV(0.05,10)	返回χ^2分布的左尾概率的反函数（3.940）
13	CHISQ.INV.RT(probability,deg_freedom)	=CHISQ.INV.RT (0.05,10)	返回χ^2分布的右尾概率的反函数（18.307）；参见 19.4.2 节
14	CHISQ.TEST(actual_range, expected_range)	=CHISQ.TEST(A2:B4, A6:B8)	对实际频数区域 A2:B4 和理论频数区域 A6:B8 进行卡方检验求 *P* 值
15	CONFIDENCE.NORM(alpha,standard_dev,size)	=CONFIDENCE.NORM(0.05,2.5,50)	使用正态分布（检验水准 α、标准差 *S*、样本例数 *n*）返回总体平均值的置信区间宽度的一半（0.693）
16	CONFIDENCE.T(alpha,standard_dev,size)	=CONFIDENCE.T(0.05,2.5,50)	使用学生的 *t* 分布（检验水准 α、标准差 *S*、样本例数 *n*）返回总体平均值的置信区间宽度的一半（0.715）
17	CORREL(array1,array2)	=CORREL(A2:A6,B2:B6)	计算区域 A2:A6 与 B2:B6 内数据间的 Pearson 相关系数
18	COUNT(value1,value2,...)	=COUNT(A2:A8)	清点区域中数值单元格的个数（不包括空白单元格、文本、逻辑值）
19	COUNTA(value1,value2,...)	=COUNTA(A2:A8)	计算区域中非空白单元格的个数（包括文本、逻辑值）
20	COUNTBLANK(range)	=COUNTBLANK(A2: B5)	计算单元格区域中空白单元格的个数
21	COUNTIF(range,criteria)	=COUNTIF(B2:B5,">55")	计算区域 B2:B5 中的数值大于 55 的单元格个数

续表

编号	函　数	举　例	说明（最末括号内为计算结果）
22	COUNTIFS(criteria_range1,criteria1,[criteria_range2, criteria2]…)	=COUNTIFS(B2:B5,">55" ，C2:C5,"<55")	计算区域 B2:B5 中的数值大于 55 且同时 C2:C5 中的数值小于 55 的单元格个数
23	COVARIANCE.P(array1,array2)	=COVARIANCE.P(A2:A6, B2:B6)	计算总体协方差，即每对数据点的离均差乘积的平均数；公式：$\left[\sum(x-\bar{x})(y-\bar{y})\right]/n$
24	COVARIANCE.S(array1,array2)	=COVARIANCE.S(A2:A6, B2:B6)	计算样本协方差，即每对数据点的离均差乘积的平均数；公式：$\left[\sum(x-\bar{x})(y-\bar{y})\right]/(n\text{-}1)$
25	DEVSQ(number1,number2,...)	=DEVSQ(A2:A8)	计算离均差平方和公式：$\sum(x-\bar{x})^2$
26	EXPON.DIST(x,lambda, cumulative)	=EXPON.DIST(0.2,10,FALSE)	获得指数分布函数（1.353353） Cumulative=0 或 FALSE 计算概率密度函数 Cumulative=1 或 TRUE 计算累积分布函数
27	F.DIST(x,deg_freedom1,deg_freedom2,cumulative)	=F.DIST(15.21,6,4,FALSE)	计算 15.21 在分子自由度为 6、分母自由度为 4 的 F 分布中的概率密度函数值（0.001）；
28	F.DIST.RT(x,deg_freedom1,deg_freedom2)	=F.DIST.RT(15.21,6,4)	计算 15.21 在分子自由度为 6、分母自由度为 4 的 *F* 分布中的 *P* 值（0.001）；参见 19.3.2 节
29	F.INV(probability,degrees_freedom1,degrees_freedom2)	=F.INV(0.01,6,4)	返回分子自由度为 7，分母自由度为 4 的 F 概率分布在其概率为 0.01 时的反函数(0.109)
30	F.INV.RT(probability,deg_freedom1,deg_freedom2)	=F.INV.RT(0.01,6,4)	返回分子自由度为 7，分母自由度为 4 的（右尾）F 概率分布在其概率为 0.01 时的反函数（15.206）
31	F.TEST(array1,array2)	=F.TEST(A2:A6, B2:B6)	判断样本 A2:A6 与 B2:B6 的方差是否不同，返回单尾概率
32	FISHER(x)	=FISHER(0.75)	相关系数为 *r*=0.75 的正态变换值（0.972955）；可采用变换后的值按正态方法求相关系数置信区间 公式：*z*＝ATANH(*r*)=0.5ln[(1+*r*)/(1-*r*)]
33	FISHERINV(y)	=FISHERINV (0.972955)	将变换值 *z*=0.972955 还原为相关系数（0.75），FISHER 的逆函数 公式：*r*=TANH(*z*)=$(e^{2y}-1)/(e^{2y}+1)$
34	FORECAST(x,known_y's, known_x's)	=FORECAST(30,A2:A6,B2:B6)	建立变量 *X*（A2:A6）与 *Y*（B2:B6）的直线回归方程，再用方程预测 *X*＝30 时对应的 *Y* 值

续表

编号	函　数	举　例	说明（最末括号内为计算结果）
35	FREQUENCY(data_array, bins_array)	=FREQUENCY(A2:A10,B2:B5)	获得连续型变量的频数表；A2:A10 为原数据，B2:B5 为频数表各个组段的上限值；参见 7.3.1 节
36	GAMMA.DIST(x,alpha,beta, cumulative)	=GAMMADIST(10,9,2,FALSE)	计算伽玛分布函数值（0.0326） Cumulative=0 或 FALSE 计算概率密度函数 Cumulative=1 或 TRUE 计算累积分布函数
37	GAMMA.INV(probability, alpha,beta）	=GAMMAINV(A2,A3,A4)	伽玛累积分布函数的逆函数（分位数）
38	GAMMALN(x)	=GAMMALN(4)	4 的伽玛函数的自然对数（1.791759）
39	GAMMALN.PRECISE(x)	=GAMMALN.PRECISE (4)	4 的伽玛函数的自然对数（1.791759）
40	GEOMEAN(number1,number2,...)	=GEOMEAN(4,5,8,7,11,4,3)	计算数据的几何均数(5.476987)；参见 7.3.2 节
41	GROWTH(known_y's, known_x's,new_x's,const)	=GROWTH(B2:B7,A2:A7, A9:A10)	根据 Y（B2:B7）和 X（A2:A7）拟合指数回归曲线，根据曲线对给定的 X 值预测 Y 值 公式：$Y=a\times b^X$（const＝TRUE（默认））$Y=b^X$（const＝FALSE）
42	HARMEAN(number1,number2,...)	=HARMEAN(A2:A8)	计算数据区域 A2:A8 的调和平均值（5.028376）；参见 7.3.2 节
43	HYPGEOM.DIST(sample_s,number_sample,population_s,number_population)	=HYPGEOM.DIST(1, 4,8,20)	获得总体成功数为 8、总体含量为 20 的超几何分布中样本成功数为 1、样本含量为 4 的概率（0.363261）
44	INTERCEPT(known_y's, known_x's)	=INTERCEPT(A2:A6, B2:B6)	利用变量 X（A2:A6）与 Y（B2:B6），计算直线回归方程的截距项
45	KURT(number1,number2,...)	=KURT(A2:A11)	求区域数据 A2:A11 的峰度系数。参见 7.3.3 节
46	LARGE(array,k)	=LARGE(A2:B6,7)	求区域 A2:B6 中第 7 大值
47	LINEST(known_y's, known_x's, const,stats)	=LINEST(A2:A5,B2:B5,,FALSE)	获得（多重）线性回归的参数；参见第 14 章；const＝TRUE（默认），按正常计算 a。const=FALSE，a＝0，方程为 $Y=bX$；stats 为 TRUE，则返回回归系数、系数标准误、决定系数、剩余标准差、方差分析 F 值、df、回归平方和、剩余平方和；stats 为 FALSE（默认），只返回常数项 a 和回归系数 b

续表

编号	函　数	举　例	说明（最末括号内为计算结果）
48	LOGEST(known_y's, known_x's,const,stats)	=LOGEST(B2:B7, A2:A7, TRUE, TRUE)	获得（多重）指数回归的参数；const 与 stats 的意义参见 LINEST 函数
49	LOGNORM.INV(probability,mean, standard_dev)	=LOGNORM.INV(0.039084, 3.5, 1.2)	求均数 3.5、标准差 1.2 的对数正态分布的 0.039084 界值（4.000028），Mean 和 Standard_dev 均为 ln(*x*)的均数和标准差
50	LOGNORM.DIST(x,mean, standard_dev)	=LOGNORM.DIST(A2, A3,A4)	对数正态分布的累积分布函数值（0.039084），参见 NORMDIST 函数
51	MAX(number1,number2,...)	=MAX(A2:A6)	求区域 A2:A6 中最大值
52	MAXA(value1,value2,...)	=MAXA(0,0.2,0.5, 0.4,TRUE)	求一组数据的最大值；包括文本和逻辑值（TRUE 为 1，文本或 FALSE 为 0）（1）
53	MEDIAN(number1,number2, ...)	=MEDIAN(A2:A7)	求数据 A2:A7 的中位数，参见 7.3.2 节
54	MIN(number1,number2,...)	=MIN(A2:A6)	求数据 A2:A6 的最小值
55	MINA(value1,value2,...)	=MINA(0,0.2,0.5, 0.4,FALSE)	求最小值；包括文本和逻辑值（FALSE 值为 0）（0）
56	MODE.MULT((number1,[number 2],...])	=MODE.MULT(A2:A7)	返回 A2:A7 区域中出现频率最高或重复出现的数值的垂直数组
57	MODE.SNGL(number1,number2,.. .)	=MODE.SNGL(A2:A7)	求数据 A2:A7 的众数，即出现频率最多的数
58	NEGBINOM.DIST(number_f, number_s,probability_s)	=NEGBINOM.DIST(10,5,0.25)	求成功概率为 0.25 的负二项式分布中达到 5 次成功前失败 10 次的概率值（0.055049）
59	NORM.DIST(x,mean,standard_de v,cumulative)	=NORM.DIST(42,40, 1.5,TRUE)	计算 42 在正态分布（40,1.5^2）中的累积分布函数（0.908789）；参见 17.1.2 节。Cumulative=0 或 FALSE 计算概率密度函数 Cumulative=1 或 TRUE 计算累积分布函数
60	NORM.INV(probability,mean,stan dard_dev)	=NORM.INV(0.908789,40,1.5)	获得正态分布累积函数的逆函数值（或分位数）（42）
61	NORM.S.DIST(z)	=NORM.S.DIST(1.645)	获得 1.645 的标准正态分布累积函数值，即 *z* 值对应的概率
62	NORM.S.INV(probability)	=NORM.S.INV(0.95)	获得标准正态累积分布函数的逆函数值；即由 *P* 值获得标准正态分布临界值（1.645）
63	PEARSON(array1,array2)	=PEARSON(A2:A6, B2:B6)	获得 Pearson 相关系数，等价于 CORREL 函数

续表

编号	函　数	举　例	说明（最末括号内为计算结果）
64	PERCENTILE.EXC(array,k)	=PERCENTILE.EXC(A2:A5,0.3)	求单元格区域 A2:A5 数据的第 30 百分位数值。此处的百分点值的范围为 0 到 1（不含 0 和 1）
65	PERCENTILE.INC(array,k)	=PERCENTILE.INC(A2:A5,0.3)	求单元格区域 A2:A5 数据的第 30 百分位数值。此处的百分比值的范围为 0 到 1（含 0 和 1）
66	PERCENTRANK.EXC(array,x,significance)	=PERCENTRANK.EXC(A2:A11, 2)	求 2 在单元格区域 A2:A11 数据中的百分比排位（如果数据中小于 2 的值有 3 个，而大于 2 的值有 6 个，则 3/(3+6)=0.333）。此处的百分点值的范围为 0 到 1（不含 0 和 1）
67	PERCENTRANK.INC(array,x,significance)	=PERCENTRANK.INC(A2:A11, 2)	求 2 在单元格区域 A2:A11 数据中的百分比排位（如果数据中小于 2 的值有 3 个，而大于 2 的值有 6 个，则 3/(3+6)=0.333）。此处的百分比值的范围为 0 到 1（含 0 和 1）
68	PERMUT(number,number_chosen)	=PERMUT(100,3)	从 100 个对象中选取 3 个对象的排列数（970200）；参见 2.9.1 节
69	POISSON.DIST(x,mean,cumulative)	=POISSON.DIST(2,5, TRUE)	均值 5 的泊松分布中发生 2 次的概率（0.124652） Cumulative=0 或 FALSE 计算概率质量函数 Cumulative=1 或 TRUE 计算累积分布函数 参见 17.3.2 节
70	PROB(x_range,prob_range,lower_limit,upper_limit)	=PROB(A2:A5, B2:B5,1,3)	计算对应概率值为 B2:B5（所有概率之和为 1）的 X 取值（A2:A5）中数值落在 1～3 区间内的概率
71	QUARTILE.EXC(array,quart)	=QUARTILE.EXC(A2:A9,1)	计算区域 A2:A9 中的第一个四分位数（第 25 百分位数）（quart＝0，1，2，3，4 分别对应第 0、25、50、75、100 百分位数）。此处的百分比值的范围为 0 到 1（不含 0 和 1）
72	QUARTILE.INC(array,quart)	=QUARTILE.INC(A2:A9,1)	计算区域 A2:A9 中的第一个四分位数（第 25 百分位数）（quart＝0，1，2，3，4 分别对应第 0、25、50、75、100 百分位数）。此处的百分比值的范围为 0 到 1（含 0 和 1）

续表

编号	函　　数	举　　例	说明（最末括号内为计算结果）
73	RANK.AVG(number,ref,order)	=RANK.AVG(A3,A2:A6,1)	A3 中数据在区域 A2:A6 中的大小秩次；相同值排秩结果以其平均秩次作为它们的秩次值；order＝1 排升序，order＝0 排降序
74	RANK.EQ(number,ref,order)	=RANK.EQ(A3,A2:A6,1)	A3 中数据在区域 A2:A6 中的大小秩次；相同值排秩结果以前次的秩次作为它们的秩次值；order＝1 排升序，order＝0 排降序
75	RSQ(known_y's,known_x's)	=RSQ(A2:A8,B2:B8)	计算 Pearson 相关系数的平方，即决定系数
76	SKEW(number1,number2,...)	=SKEW(A2:A11)	获得区域 A2:A11 数据的偏度系数
77	SLOPE(known_y's,known_x's)	=SLOPE(A2:A8, B2:B8)	用变量 Y（A2:A8）与 X（B2:B8）计算直线回归方程的斜率（或回归系数）
78	SMALL(array,k)	=SMALL(B2:B10,2)	求区域 B2:B10 中第 2 较小值
79	STANDARDIZE(x,mean, standard_dev)	=STANDARDIZE(42, 40,1.5)	将一般正态分布（μ,σ^2）转换为标准正态分布（0，1^2）（1.333333） 公式：$z=(X-\bar{X})/S$
80	STDEV.P(number1,number2,...)	=STDEV.P(A2:A11)	计算基于以参数形式给出的整个样本总体的标准偏差 公式：$\sqrt{\frac{\sum(x-\bar{x})^2}{n}}$
81	STDEV.S(number1,number2,...)	=STDEV.S(A2:A11)	基于样本估算标准偏差 公式：$\sqrt{\frac{\sum(x-\bar{x})^2}{(n-1)}}$
82	STDEVA(value1,value2,...)	=STDEVA(A2:A11)	基于样本（包括数字、文本和逻辑值）估算标准偏差 公式：$\sqrt{\frac{\sum(x-\bar{x})^2}{(n-1)}}$
83	STDEVPA(value1,value2,...)	=STDEVPA(A2:A11)	基于样本总体（包括数字、文本和逻辑值）计算标准偏差 公式：$\sqrt{\frac{\sum(x-\bar{x})^2}{n}}$
84	STEYX(known_y's,known_x's)	=STEYX(A2:A8, B2:B8)	计算线性回归方程的剩余标准差 公式：$S_{y.x}=\sqrt{\frac{\Sigma(Y-\hat{Y})^2}{n-2}}=\sqrt{\frac{SS_{剩}}{n-2}}$
85	T.DIST(x,deg_freedom, cumulative)	=T.DIST(1.96,60,FALSE)	返回学生的 t 分布的概率密度函数(0.0598)；参见 19.2.2 节

续表

编号	函　数	举　例	说明（最末括号内为计算结果）
86	T.DIST.2T(x,deg_freedom)	=T.DIST.2T(1.96,60)	返回学生的双尾 t 分布(0.0546)
87	T.DIST.RT(x,deg_freedom)	=T.DIST.RT(1.96,60)	返回学生的右尾 t 分布(0.0273)
88	T.INV(probability, degrees_freedom)	=T.INV(0.054645, 60)	返回学生的 t 分布的左尾反函数（-1.626）
89	T.INV.2T(probability,degrees_freedom)	=T.INV.2T(0.054645, 60)	返回学生 t 分布的双尾反函数（1.960）
90	TREND(known_y's,known_x's, new_x's,const)	=TREND(B2:B13, A2:A13,A15:A19)	由变量 Y（B2:B13（与 X（A2:A13（建立回归方程，获得 X（A15:A19）在直线上的预测值
91	TRIMMEAN(array,percent)	=TRIMMEAN(A2:A12,0.2)	去掉两端 20%的极端值后，计算单元格区域 A2:A12 数据的修整均数；参见 7.3.2 节
92	T.TEST(array1,array2,tails,type)	=T.TEST(A2:A10, B2:B10,2,1)	计算 t 检验结果，type＝1、2、3 分别对应配对、等方差和异方差双样本）；tails 同前
93	VAR.S(number1,number2,...)	=VAR.S(A2:A11)	估计区域 A2:A11 的样本方差（即标准差的平方）；参见 7.3.3 节 公式：$S^2=\sum_{i=1}^{n}\frac{(X_i-\bar{X})^2}{n-1}$
94	VARA(value1,value2,...)	=VARA(A2:A11)	计算样本方差，包括文本值和逻辑值
95	VAR.P(number1,number2,...)	=VAR.P(A2:A11)	估计区域 A2:A11 的总体方差 公式：$\sigma^2=\sum_{i=1}^{n}\frac{(X_i-\mu)^2}{n}$
96	VARPA(value1,value2,...)	=VARPA(A2:A11)	估计总体方差，包括文本值和逻辑值
97	WEIBULL.DIST(x,alpha,beta, cumulative)	=WEIBULL.DIST(A2,A3,A4,TRUE)	返回 Weibull 分布的累积函数 Cumulative 同前
98	Z.TEST(array,x,sigma)	=Z.TEST(A2:A11,4)	返回总体均数＝4 相对于区域 A2:A11 的 z 检验（左尾）单侧 P 值。用于单样本与总体均数的比较。sigma 为总体标准差，省略则使用样本标准差替代 公式：$P=1-\text{NORMSDIST}\left(\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}\right)$

3. 日期与时间函数

（1）日期函数（date functions）

编号	函　数	举　例	说明（最末括号内为计算结果）
1	DATE(year,month,day)	=DATE(2000,1,1)	Excel 规定 1899 年 12 月 31 日为 0 点，1900 年 1 月 1 日的序列号是 1，2000 年 1 月 1 日与 1900 年 1 月 1 日之间包括了 36526 天。因此该函数产生日

续表

编号	函　数	举　例	说明（最末括号内为计算结果）
			期（常规格式结果为 2000-1-1 或序列数值格式结果为 36526）
2	DATEDIF(start_date, end_date,"y")	=DATEDIF(A1,TODAY(),"y")	计算 A1 日期到今天的年数；将 y 改为 m，则计算两个日期之间的月数；将 y 改为 d，则计算两个日期之间的天数；如果 A1 是 2000-1-1，今天是 2007-12-2，则得到 7 年；将 y 改为 m 则得到 95 月；将 y 改为 d 则得到 2892 天
3	DATEVALUE(date_text)	=DATEVALUE("2000-1-1")	将文本日期转换为序列数日期（36526）
4	DAY(serial_number)	=DAY(A1)	寻找日期是一个月的第几天（取值 1～31）。如果 A1=2007-12-2，则获得结果为 2
5	DAYS360(start_date, end_date,method)	=DAYS360(A1,B1)	按照一年 360 天的算法，计算 A1、B1 两个日期之间的天数
6	EDATE(start_date, months)	=EDATE(A2,-2) =EDATE(A2,2)	返回 A2 日期之前（后）2 个月的日期（或序列数），如果 A2＝2007-12-2，则得到前者为 2007-10-2（或 39357），后者为 2008-2-2（或 39480） 安装了数据分析工具后可用
7	EOMONTH(start_date,months)	=EOMONTH(A2,2) =EOMONTH(A2,3)	返回 A2 日期之后 2（3）个月的最后一个日期（或序列数），如果 A2＝2007-12-2，则得到前者为 2008-2-29（或 39507），后者为 2008-3-31（或 39538） 安装了数据分析工具后可用
8	MONTH(serial_number)	=MONTH("2007-12-2")	返回日期值的月份，取值 1（一月）到 12（十二月）的整数（12）
9	NETWORKDAYS(start_date, end_date,holidays)	=NETWORKDAYS("2002/05/01","2002/05/31","2002/05/28")	返回"2002/05/01"到"2002/05/31"之间的工作日数，不包括周末和所列出假日"2002/05/28"（22） 安装了数据分析工具后可用
10	NETWORKDAYS.INTL(start_date,end_date, [weekend], [holidays])	=NETWORKDAYS.INTL("2002/05/01","2002/05/31",7,"2002/05/28")	返回"2002/05/01"到"2002/05/31"之间的工作日数，不包括周末（指定为星期五和星期六为周末）和所列出假日"2002/05/28"（22） 安装了数据分析工具后可用
11	TODAY()	=TODAY()	返回当前日期的序列数（2007-12-2）
12	WEEKDAY(serial_number, return_type)	=WEEKDAY(A2,2) =WEEKDAY("2007-12-2",2)	返回 A2 的星期号，数字 1～7 分别代表星期一～星期日。这里“return_type”有 1（默认）、2、3 三种选择，如果选择 1，结果 1～7 分别代表星期日～星期六；选择 2，结果 1～7 分别代表星期一～星期天；选择 3，则 0～6 代表星期日～星期六。本例 return_type=2，得到结果为星期天（7）

续表

编号	函　数	举　例	说明（最末括号内为计算结果）
13	WEEKNUM(serial_num, return_type)	=WEEKNUM("2007-12-2",1)	返回日期是一年中的第几周，Return_type 为 1（默认），一周开始于星期日；为 2 则开始于星期一（49）
14	WORKDAY(start_date,days, holidays)	=WORKDAY(A1,151, A4:A6)	从日期 A1 开始隔 151 个工作日的日期值，除去假日单元格区域 A4:A6 的日期。如果 A1=2007-1-1，A4:A6 无数据，则得到 2007-7-31 或序列数据值 39294 安装了数据分析工具后可用
15	WORKDAY.INTL(start_date, days, [weekend], [holidays])	=WORKDAY(A1,151, 7,A4:A6)	从日期 A1 开始隔 151 个工作日（指定为星期五和星期六为周末）的日期值，除去假日单元格区域 A4:A6 的日期。如果 A1=2007-1-1，A4:A6 无数据，则得到 2007-7-31 或序列数据值 39294 安装了数据分析工具后可用
16	YEAR(serial_number)	=YEAR("1998-7-5")	返回日期 1998-7-5 的年份数，取值 1900～9999 的整数（1998）
17	YEARFRAC(start_date, end_date,basis)	=YEARFRAC(A2,A3,3)	日期 A2 与 A3 之间的天数占全年的百分比

（2）时间函数（time functions）

编号	函　数	举　例	说明（最末括号内为计算结果）
1	HOUR(serial_number)	=HOUR(0.7)	返回时间值的小时数，取值 0（12:00A.M.）到 23（11:00P.M.）的整数（16）
2	MINUTE(serial_number)	=MINUTE("4:48:00 PM")	返回时间值的分钟数，取值 0 到 59 的整数（48）
3	NOW()	=NOW()	返回当前日期和时间或序列数，如当前时间为 2007-12-2　16:31:02（或 39418.6882）
4	SECOND(serial_number)	=SECOND("4:48:18 PM")	返回时间值的秒数，取值 0 到 59 的整数（18）
5	TIME(hour,minute,second)	=TIME(16,48,10)	返回时间小数值，取值 0 到 0.9999，代表 0:00:00 到 23:59:59 的时间（0.700115741）
6	TIMEVALUE(time_text)	=TIMEVALUE("2:24 AM")	返回文本型时间的小数值，同上（0.1）

4. 文本函数（text functions）

编号	函　数	举　例	说明（最末括号内为计算结果）
1	ASC(text)	=ASC("ＥＸＣＥＬ")	将全角英文字符更改为半角字符（EXCEL）
2	BAHTTEXT(number)	=BAHTTEXT(65)	将数字转换为泰铢文字（หกสิบห้าบาท ทถ้วน）

续表

编号	函　　数	举　　例	说明（最末括号内为计算结果）
3	CHAR(number)	=CHAR(65)	返回对应于 ANSI 数字代码的字符（"A"）
4	CLEAN(text)	=CLEAN(CHAR(7)&"text")	删除文本中不能打印的字符（text）
5	CODE(text)	=CODE("Alphabet")	返回文字串中第 1 个字符"A"的数字代码（65）
6	CONCATENATE (text1,text2,...)	=CONCATENATE(A2," ",A3," ")	将几个字符串合并为一个字符串。等价于&（和号）运算符
7	DOLLAR (number,decimals)	=DOLLAR(1234.567, -2)	以货币格式显示，保留小数点前 2 位“$1200”
8	EXACT(text1,text2)	=EXACT(A2,B2)	测试字符串 A2 与 B2 是否完全相同。相同则为 TRUE；否则为 FALSE
9	FIND(find_text, within_text,start_num)	=FIND("M",A2)	查找字符串 A2 中“M”的位置（区分大小写并且不允许使用通配符）；FINDB 用于双字节字符
10	FIXED(number,decimals,no_commas)	=FIXED(-1234.56,-1)	将数值转换为字符串形式，保留小数点前 1 位（"-1230"）；No_commas 为 TRUE，字符串不含逗号；为 FALSE（默认），则包含逗号
11	JIS(text)	=JIS("EXCEL")	将半角英文字符更改为全角字符("ＥＸＣＥＬ")，仅日文版中有效
12	LEFT(text,num_chars)	=LEFT(A2,4)	返回字符串 A2 中的前 4 个字符；LEFTB 用于双字节字符
13	LEN(text)	=LEN("Phoenix,AZ")	获得字符串的长度（10） LENB 用于双字节字符
14	LOWER(text)	=LOWER("Apt. 2B")	将字符串转换为小写（"apt. 2b"）
15	MID(text,start_num, num_chars)	=MID("FluidFlow",1,5)	返回字符串中第 1 个字符开始的 5 个字符（Fluid）；MINB 用于双字节字符
16	PHONETIC(reference)	=PHONETIC(C4)	从字符串 C4 中提取拼音字符，仅日文版中有效
17	PROPER(text)	=PROPER("76BGet")	将首字母及任何非字母字符后的首字母转换成大写，其余字母转换成小写（76Bget）
18	REPLACE(old_text,start_num , num_chars, new_text)	=REPLACE("abcdefghijk",6, 5, "*")	从字符串第 6 个字符开始的连续 5 个字符用"*"替换（abcde*k） REPLACEB 用于双字节字符
19	REPT(text,number_times)	=REPT("*-", 3)	重复显示字符串"*-"3 次（*-*-*-）
20	RIGHT(text,num_chars)	=RIGHT("Sale Price",5)	返回字符串的最后 5 个字符（"Price"）
21	RMB(number,decimals)	=RMB(12.567, 2))	以货币格式显示，保留 2 位小数（￥12.57）
22	SEARCH(find_text,within_text,start_num)	=SEARCH("e","Statements", 6)	返回从字符串第 6 个字符开始首次出现"e"的位置（7）；SEARCHB 用于双字节字符

续表

编号	函　　数	举　　例	说明（最末括号内为计算结果）
23	SUBSTITUTE(text,old,new,[num])	=SUBSTITUTE("Quarter 1, 1991", "1", "2", 1)	在字符串"Quarter 1, 1991"中用"2"替换第 1 个"1"（Quarter 2, 1991） num 缺省则替换全部，本例 num＝1
24	T(value)	=T(253333a) =T(253333)	文本则返回文本；数字则返回空文本（""） 前者得到 253333a，而后者格子无任何内容显示
25	TEXT(value,format_text)	=TEXT(2.715, "$0.00")	将数值转换为按指定数字格式表示的文本（$2.72）
26	TRIM(text)	=TRIM("First Quarter Earnings")	清除文本中除单词之间的单个空格外的所有空格（First Quarter Earnings）
27	UPPER(text)	=UPPER("total")	将文本转换成大写形式（"TOTAL"）
28	VALUE(text)	=VALUE("$1,000")	将代表数字的文字串转换成数字（1000）
29	WIDECHAR(text)	=WIDECHAR("EXCEL")	将字符串中的单字节字符转换为双字节字符（ＥＸＣＥＬ）
30	YEN(number,decimals)	=YEN(-0.123,4)	用￥（日圆）货币格式将数字转换成文本（￥-0.1230），仅日文版中有效

5. 逻辑函数（logical functions）

编号	函　　数	举　　例	说明（最末括号内为计算结果）
1	AND(logical1,logical2, ...)	=AND(1+1=2, 2+2=5)	所有各项均正确则返回 TRUE；只要有一个为假则返回 FALSE（FALSE）
2	FALSE()		返回逻辑值 FALSE，也可在工作表或公式中直接输入
3	IF(logical_test,value_if_true,value_if_false)	=IF(A5=100,SUM(B5:B25),"")	执行真假值判断，根据逻辑测试的真假值返回不同的结果。如果 A5=100 为真，则计算区域 B5:B25 的和，否则返回空文本
4	IFERROR(value, value_if_error)	=IFERROR(A2/B2," 计 算中有错误")	如果公式的计算结果为错误，则返回"计算中有错误"；否则将返回公式的结果
5	NOT(logical)	=NOT(1+1=2) =NOT(1+1=3)	对逻辑公式求反 前者得到（FALSE），而后者得到 TRUE
6	OR(logical1,logical2,...)	=OR(1+1=2, 2+2=5)	任何一项正确则返回 TRUE，否则为假（TRUE），即全错才返回 FALSE
7	TRUE()		返回逻辑值 TRUE，也可在工作表或公式中直接输入

6. 查找和引用函数

（1）查找函数（lookup functions）

编号	函　　数	举　　例	说明（最末括号内为计算结果）
1	CHOOSE(num,value1, value2,...)	=CHOOSE(2,"1st","2nd","3rd")	返回 3 个参数值中的第 2 个参数值（"2nd"）
2	GETPIVOTDATA(data,table,field1,item1,...)	GETPIVOTDATA("求和项：销售额",$A$4,"月份","三月")	提取存储在数据透视表中的数据（由 A4 指定所需提取的数据） 返回“三月”的总计值
3	HLOOKUP(lookup_value, table_array,row_index_num,range_lookup)	=HLOOKUP(2,{1,2,3;"a","b","c";"d","e","f"},3,TRUE)	在说明的行内查找值。在数组常量（见公式中花括号中的内容）中查找 2，并在该值所在列中的第 3 行查找值（e）
4	①INDEX(array,row_num,column_num) ②INDEX(reference,row_num,column_num,area_num)	①=INDEX({1,2;3,4},2,1) ②=INDEX((A1:C6, A8:C11),2,2,1)	①返回数组中的第 2 行第 1 列的元素值（3） ②返回引用中第 1 个区域的第 2 行第 2 列的元素值（B2）
5	①LOOKUP(lookup_value,lookup_vector,result_vector) ②LOOKUP(lookup_value,array)	①=LOOKUP(4.9,A2:A7,B2:B7) ②=LOOKUP("C",{"a",1,4;"b",2,5;"c",3,6})	①在 A2:A7 中查找数值 4.9，返回 B2:B7 中相同位置的数值。注意区域只包含一行或一列（向量）。如果 A4=4.91，B4=9，则（9） ②在数组的第一行（列数＞行数）或第一列（列数≤行数）中查找"C"，然后返回最后一行或最后一列中相同位置处的数值（6）
6	MATCH(lookup_value,lookup_array,match_type)	=MATCH("b",{"a","b","c"},0)	返回与"b"匹配的数组中元素的相应位置（2）。如果 match_type 为 1（默认），查找≤lookup_value 的最大数值，lookup_array 必须升序排列；为 0，查找= lookup_value 的第一个数值；为-1，查找≥lookup_value 的最小数值，lookup_array 必须降序排列
7	RTD(progID,server,topic1,[topic2,...])	略	支持 com 自动化的程序中提取数据
8	VLOOKUP(lookup_value,table_array,col_index_num,range_lookup)	=VLOOKUP(2,{1,2,3;"a","b","c";"d","e","f"},3,TRUE)	在说明的列内查找值。在数组常量（见公式中花括号中的内容）中查找 2，并在该值所在行中的第 3 列查找值（3）

（2）引用函数（reference functions）

编号	函　　数	举　　例	说明（最末括号内为计算结果）
1	ADDRESS(row_num, column_num,abs_num, a1,sheet_text)	=ADDRESS(2,3,2)	以文字形式返回对工作簿中某一单元格的引用。按行号 2 和列标 3，绝对行号相对列标建立文本类型的单元格地址（"C$2"）。abs_num 取值 1～4，分别对应绝对引用（默认）、绝对行号相对列标、相对行号绝对列标、相对引用；a1 为 TRUE 或 FALSE，分别返回 A1 或 R1C1 样式
2	AREAS(reference)	=AREAS((B2:D4,E5,F6:I9))	返回引用中包含的区域个数（3）
3	COLUMN(reference)	=COLUMN(B3) =COLUMN(C3:D5)	返回引用单元格的列标，A、B、…分别记为 1、2、…（2）（3） 其中 COLUMN()表示引用所在活动单元格
4	COLUMNS(array)	=COLUMNS(A1:C4) =COLUMNS({"a",1,4;"b",2,5;"c",3,6})	返回数组或引用的列数（3）（3）
5	HYPERLINK(link_location, friendly_name)	=HYPERLINK("[Budget]June!E56", "Click here")	创建超级链接。创建跳转到 Budget 工作簿 June 工作表中单元格 E56 的超级链接，链接的文本为"Click here"
6	INDIRECT(ref_text,a1)	=INDIRECT($A$1)	返回$A$1 中指定的引用。如果单元格 A1 包含文本"B2"，B2 包含数值 1.33，则（1.33）
7	OFFSET(reference,rows, cols,height,width)	=SUM(OFFSET(C3:E5,-1,1,4,4))	以"reference"为参照，向下偏"rows"行，向右偏"cols"列。返回一个"height"行、"width"列的引用区域。本例返回指定引用 C3:E5 向下偏移-1 行（即向上移一行）、向右偏移 1 列的 4 行 4 列引用区域（D2:G5）。如果 D2:G5 区域之和恰好为 0，则（0）
8	ROW(reference)	=ROW(B3) = ROW(C3:D5)	返回引用单元格的行号，1、2、……分别记为 1、2、……（3）（3） 其中 ROW()表示引用所在活动单元格
9	ROWS(array)	=ROWS({1,2,3;4,5,6})	返回引用或数组的行数（2）
10	TRANSPOSE(array)	=TRANSPOSE($A$1:$C$1)	将数组行列互换即转置。函数必须在某个区域中以数组公式形式输入，该区域行、列数分别为数组的列、行数

7. 数据库函数（database functions）

编号	函　数	举　例	说明（最末括号内为计算结果）
1	DAVERAGE(database,field,criteria)	=DAVERAGE(A5:E11,3, A1:B2)	返回数据库区域 A5:E11 中满足 A1:B2 中给定条件的第 3 列中数值的平均值
2	DCOUNT(database,field,criteria)	=DCOUNT(A5:E11,3,A1:B2)	返回数据库区域 A5:E11 中满足 A1:B2 中给定条件的第 3 列中含数字的单元格数目
3	DCOUNTA(database,field,criteria)	=DCOUNTA(A5:E11,"树龄",A1:B2)	返回数据库区域 A5:E11 中"树龄"字段里满足 A1:B2 中给定条件的非空单元格数目
4	DGET(database,field,criteria)	=DGET(A5:E11,"产量", A1:B2)	返回数据库区域 A5:E11 中"产量"字段里满足 A1:B2 中给定条件且唯一存在的值
5	DMAX(database,field,criteria)	=DMAX(A5:E11, "利润",A1:A3)	返回区域 A5:E11 的"利润"列中，满足 A1:A3 中给定条件的单元格中的最大数值
6	DMIN(database,field,criteria)	=DMIN(A5:E11, "利润",A1:A3)	返回区域 A5:E11 的"利润"列中，满足 A1:A3 中给定条件的单元格中的最小数值
7	DPRODUCT(database,field,criteria)	=DPRODUCT(A5:E11, "产量",A1:A3)	返回区域 A5:E11 的"产量"列中，满足 A1:A3 中给定条件的单元格中数值乘积
8	DSTDEV(database,field,criteria)	=DSTDEV(A5:E11, "产量",A1:A3)	返回 A5:E11 的"产量"列中，满足 A1:A3 中给定条件的单元格中数字作为一个样本，估算样本标准差
9	DSTDEVP(database,field,criteria)	=DSTDEVP(A5:E11, "产量",A1:A3)	返回 A5:E11 的"产量"列中，满足 A1:A3 中给定条件的单元格中数字作为样本，估算对应的总体标准差
10	DSUM(database,field,criteria)	=DVAR(A5:E11,"产量",A1:A3)	返回 A5:E11 的"产量"列中，满足 A1:A3 中给定条件的单元格中数字之和
11	DVAR(database,field,criteria)	=DVAR(A5:E11,"产量",A1:A3)	返回 A5:E11 的"产量"列中，满足 A1:A3 中给定条件的单元格中数字作为一个样本，估算样本方差
12	DVARP(database,field,criteria)	=DVARP(A5:E11, "产量",A1:A3)	返回 A5:E11 的"产量"列中，满足 A1:A3 中给定条件的单元格中数字作为样本，估算总体方差
13	GETPIVOTDATA(pivot_table,name)	=GETPIVOTDATA(PT1,"销售总额")	返回存储在数据透视表报表中的数据

8. 财务函数（financial functions）

编号	函　数	用　途
1	ACCRINT(issue,first_interest, settlement,rate,par,frequency,basis)	返回定期付息有价证券的应计利息

续表

编号	函　数	用　途
2	ACCRINTM(issue,maturity,rate,par,basis)	返回到期一次性付息有价证券的应计利息
3	AMORDEGRC(cost,date_purchased,first_period,salvage,period,rate,basis)	返回每个会计期间的折旧值
4	MORLINC(cost,date_purchased,first_period,salvage,period,rate,basis)	返回每个会计期间的折旧值，该函数为法国会计系统提供。如果某项资产是在会计期间内购入的，则按线性折旧法计算
5	COUPDAYBS(settlement,maturity,frequency,basis)	返回当前付息期内截止到成交日的天数
6	COUPDAYS(settlement，maturity，frequency，basis)	返回成交日所在的付息期的天数
7	COUPDAYSNC(settlement,maturity,frequency,basis)	返回从成交日到下一付息日之间的天数
8	COUPNCD(settlement,maturity,frequency,basis)	返回成交日之后的的下一个付息日期
9	COUPNUM(settlement,maturity,frequency,basis)	返回成交日和到期日之间的利息应付次数，向上取整到最近的整数
10	COUPPCD(settlement,maturity,frequency,basis)	返回成交日之前的上一付息日的日期
11	CUMIPMT(rate,nper,pv,start_period,end_period,type)	返回一笔贷款在给定的 start-period 到 end-period 期间累计偿还的利息数额
12	CUMPRINC(rate,nper,pv,start_period,end_period,type)	返回一笔贷款在给定的 start-period 到 end-period 期间累计偿还的本金数额
13	DB(cost,salvage,life,period,month)	使用固定余额递减法，计算一笔资产在给定期间内的折旧值
14	DDB(cost,salvage,life,period,factor)	使用双倍余额递减法或其他指定方法，计算一笔资产在给定期间内的折旧值
15	DISC(settlement,maturity,pr,redemption,basis)	返回有价证券的贴现率
16	DOLLARDE(fractional_dollar,fraction)	将按分数表示的价格转换为按小数表示的价格，如证券价格，转换为小数表示的数字
17	DOLLARFR(decimal_dollar,fraction)	将按小数表示的价格转换为按分数表示的价格
18	DURATION(settlement,maturity,couponyld,frequency ，basis	返回假设面值$100 的定期付息有价证券的修正期限。期限定义为一系列现金流现值的加权平均值，用于计量债券价格对于收益率变化的敏感程度
19	EFFECT(nominal_rate,npery)	利用给定的名义年利率和一年中的复利期次，计算实际年利率
20	FV(rate,nper,pmt,pv,type)	基于固定利率及等额分期付款方式，返回某项投资的未来值
21	FVSCHEDULE(principal,schedule)	基于一系列复利返回本金的未来值，用于计算某项投资在变动或可调利率下的未来值

续表

编号	函　　数	用　　途
22	INTRATE(settlement,maturity,investment,redemption ，basis)	返回一次性付息证券的利率
23	IPMT(rate,per,nper,pv,fv,type)	基于固定利率及等额分期付款方式，返回投资或贷款在某一给定期限内的利息偿还额
24	IRR(values,guess)	返回由数值代表的一组现金流的内部收益率
25	ISPMT(rate,per,nper,pv)	计算特定投资期内要支付的利息
26	MDURATION(settlement,maturity,coupon,yld,frequency,basis)	返回假设面值$100 的有价证券的 Macauley 修正期限
27	MIRR(values,finance_rate,reinvest_rate)	返回某一期限内现金流的修正内部收益率
28	NOMINAL(effect_rate,npery)	基于给定的实际利率和年复利期数，返回名义年利率
29	NPER(rate,pmt,pv,fv,type)	基于固定利率及等额分期付款方式，返回某项投资（或贷款）的总期数
30	NPV(rate,value1,value2,...)	通过使用贴现率以及一系列未来支出（负值）和收入（正值），返回一项投资的净现值
31	ODDFPRICE(settlement,maturity,issue,first_coupon,rate,yld,redemption,frequency,basis)	返回首期付息日不固定的面值$100 的有价证券的价格
32	ODDFYIELD(settlement,maturity,issue,first_coupon,rate,pr,redemption,frequency,basis)	返回首期付息日不固定的有价证券（长期或短期）的收益率
33	ODDLPRICE(settlement,maturity,last_interest,rate,yld,redemption,frequency,basis)	返回末期付息日不固定的面值$100 的有价证券（长期或短期）的价格
34	ODDLYIELD(settlement,maturity,last_interest,rate,pr,redemption,frequency,basis)	返回末期付息日不固定的有价证券（长期或短期）的收益率
35	PMT(rate,nper,pv,fv,type)	基于固定利率及等额分期付款方式，返回贷款的每期付款额
36	PPMT(rate,per,nper,pv,fv,type)	基于固定利率及等额分期付款方式，返回投资在某一给定期间内的本金偿还额
37	PRICE(settlement,maturity,rate,yld,redemption,frequency,basis)	返回定期付息的面值$100 的有价证券的价格
38	PRICEDISC(settlement,maturity,discount,redemption,basis)	返回折价发行的面值$100 的有价证券的价格
39	PRICEMAT(settlement,maturity,issue,rate,yld,basis)	返回到期付息的面值$100 的有价证券的价格
40	PV(rate,nper,pmt,fv,type)	返回投资的现值（即一系列未来付款的当前值的累积和），如借入方的借入款即为贷出方贷款的现值
41	RATE(nper,pmt,pv,fv,type,guess)	返回年金的各期利率。函数 RATE 通过迭代法计算得出，并且可能无解或有多个解

续表

编号	函　　数	用　　途
42	RECEIVED(settlement,maturity,investment,discount,basis)	返回一次性付息的有价证券到期收回的金额
43	SLN(cost,salvage,life)	返回某项资产在一个期间中的线性折旧值
44	SYD(cost,salvage,life,per)	返回某项资产按年限总和折旧法计算的指定期间的折旧值
45	TBILLEQ(settlement,maturity,discount)	返回国库券的等效收益率
46	TBILLPRICE(settlement,maturity,discount)	返回面值$100 的国库券的价格
47	TBILLYIELD(settlement,maturity,pr)	返回国库券的收益率
48	VDB(cost,salvage,life,start_period,end_period,factor,no_switch)	使用双倍余额递减法或其他指定的方法，返回指定的任何期间内（包括部分期间）的资产折旧值
49	XIRR(values,dates,guess)	返回一组现金流的内部收益率，这些现金流不一定定期发生。若要计算一组定期现金流的内部收益率，可以使用 IRR 函数
50	XNPV(rate,values,dates)	返回一组现金流的净现值，这些现金流不一定定期发生。若要计算一组定期现金流的净现值，可以使用函数 NPV
51	YIELD(settlement,maturity,rate,pr,redemption,frequency,basis)	返回定期付息有价证券的收益率，函数 YIELD 用于计算债券收益率
52	YIELDDISC(settlement,maturity,pr,redemption,basis)	返回折价发行的有价证券的年收益率
53	YIELDMAT(settlement,maturity,issue,rate,pr,basis)	返回到期付息的有价证券的年收益率

9. 信息函数（information functions）

编号	函　　数	用　　途
1	CELL(info_type，reference)	返回某一引用区域的左上角单元格的格式、位置或内容等信息，该函数主要用于保持与其他电子表格程序的兼容性
2	ERROR.TYPE(error_val)	返回对应于某一错误类型的数字，如果没有错误则返回#N/A。在 IF 函数中可以使用 ERROR.TYPE 检测错误值，并返回文字串（如“发生错误”）来取代错误值
3	INFO(type_text)	返回有关当前操作环境的信息
4	ISBLANK(value) 、 ISERR(value) 、 ISERROR(value)、ISLOGICAL(value)、 ISNA(value) 、 ISNONTEXT(value) 、 ISNUMBER(value) 、 ISREF(value) 、 ISTEXT(value)	其中包括用来检验数值或引用类型的 9 个工作表函数。它们可以检验数值的类型，并根据参数的值返回 TRUE 或 FALSE。例如，数值为空白单元格引用时，ISBLANK 函数返回逻辑值 TRUE，否则返回 FALSE

续表

编号	函　数	用　途
5	ISEVEN(number)	测试参数的奇偶性，如果参数为偶数，则返回 TRUE，否则返回 FALSE
6	ISODD(number)	测试参数的奇偶性，如果参数为奇数，则返回 TRUE，否则返回 FALSE
7	N(value)	返回转化为数值后的值
8	NA()	返回错误值#N/A。#N/A 表示“无法得到有效值”。如果使用#N/A 标志空白单元格，可以避免不小心将空白单元格计算在内而产生的问题，因为公式引用含有#N/A 的单元格时返回错误值#N/A
9	TYPE(value)	返回数值的类型。当某一个函数的计算结果取决于特定单元格中数值的类型时，可使用函数 TYPE；value 可以是 Excel 中的数据，如数字、文本、逻辑值等

10. 工程函数（engineering functions）

编号	函　数	用　途
1	BESSELI(x，n)	返回修正 Bessel 函数值，它与用纯虚数参数运算时的 Bessel 函数值相等
2	BESSELJ(x，n)	返回 Bessel 函数值
3	BESSELK(x，n)	返回修正 Bessel 函数值，它与用纯虚数参数运算时的 Bessel 函数值相等
4	BESSELY(x，n)	返回 Bessel 函数的值，也称为 Weber 函数或 Neumann 函数
5	BIN2DEC(number)	将二进制数转换为十进制数
6	BIN2HEX(number，places)	将二进制数转换为十六进制数
7	BIN2OCT(number，places)	将二进制数转换为八进制数
8	COMPLEX(real_num，i_num，suffix)	将实系数及虚系数转换为复数
9	CONVERT(number，from_unit，to_unit)	将数字从一个度量系统转换到另一个度量系统中
10	DEC2BIN(number，places)	将十进制数转换为二进制数
11	DEC2HEX(number，places)	将十进制数转换为十六进制数
12	DEC2OCT(number，places)	将十进制数转换为八进制数
13	DELTA(number1，number2)	测试两个数值是否相等。如果 number1=number2，则返回 1，否则返回 0
14	ERF(lower_limit，upper_limit)	返回误差函数在上、下限之间的积分
15	ERF.PRECISE(x)	返回误差函数
16	ERFC(x)	返回互补误差函数
17	ERFC.PRECISE(x)	返回从 x 到无穷大积分的互补 ERF 函数
18	GESTEP(number，step)	如果 Number 大于等于 step，则返回 1，否则返回 0。使用该函数可筛选数据

续表

编号	函　数	用　途
19	HEX2BIN(number，places)	将十六进制数转换为二进制数
20	HEX2DEC(number)	将十六进制数转换为十进制数
21	HEX2OCT(number，places)	将十六进制数转换为八进制数
22	IMABS(inumber)	返回复数的绝对值
23	IMAGINARY(inumber)	返回复数的虚系数
24	IMARGUMENT(inumber)	返回以弧度表示的角
25	IMCONJUGATE(inumber)	返回复数的共轭复数
26	MCOS(inumber)	返回复数的余弦
27	IMDIV(inumber1，inumber2)	返回两个复数的商
28	IMEXP(inumber)	返回复数的指数
29	IMLN(inumber)	返回复数的自然对数
30	IMLOG10(inumber)	返回复数的常用对数
31	IMLOG2(inumber)	返回复数的以 2 为底数的对数
32	IMPOWER(inumber，number)	返回复数的 *n* 次幂
33	IMPRODUCT(inumber1，inumber2，...)	返回 2～29 个复数的乘积
34	IMREAL(inumber)	返回复数的实系数
35	IMSIN(inumber)	返回复数的正弦值
36	IMSQRT(inumber)	返回复数的平方根
37	IMSUB(inumber1，inumber2)	返回两个复数的差
38	IMSUM(inumber1，inumber2，...)	返回两个或多个复数的和
39	OCT2BIN(number，places)	将八进制数转换为二进制数
40	OCT2DEC(number)	将八进制数转换为十进制数
41	OCT2HEX(number，places)	将八进制数转换为十六进制数

11. 多维数据集函数（cube functions）

编号	函　数	用　途
1	CUBEKPIMEMBER(connection，kpi_name, kpi_property, caption)	返回关键性能指标（KPI）属性，并在单元格中显示 KPI 名称
2	CUBEMEMBER(connection，member_express, caption)	从多维数据集返回成员或元组
3	CUBEMEMBERPROPERTY(connection, member_express, property)	从多维数据集返回成员属性的值
4	CUBERANKEDMEMBER(connection, set_express, rank, caption)	返回集合中的第 *N* 个成员

续表

编号	函　数	用　途
5	CUBESET(connection, set_express, caption, sort_order, sort_by)	通过向服务器上的多维数据集发送一组表达式来定义成员或元组的计算集，该服务期会创建集合，然后将该集合返回到微软 Excel
6	CUBESETCOUNT(set)	返回集合中分项数
7	CUBEVALUE(set)	从多维数据集返回聚合值

12. 与加载项一起安装的用户定义的函数

编号	函　数	用　途
1	CALL(register_id,[argument1],...])/(module_text,procedure,type_text,[argument1],...])	调用动态链接库或代码源中的过程
2	EUROCONVERT(number,source,target,full_precision,triangulation_precision)	用于将数字转换为欧元形式，将数字由欧元形式转换为欧元成员国货币形式，或利用欧元作为中间货币将数字由某一欧元成员国货币转化为另一欧元成员国货币形式（三角转换关系）
3	REGISTER.ID(module_text,procedure,[type_text])	返回已注册过的指定动态链接库 (DLL) 或代码源的注册号
4	SQL.REQUEST(connection_string,[output_ref],[driver_prompt],[query_text],[col_names_logical])	连接到一个外部的数据源并从工作表中运行查询，然后将查询结果以数组的形式返回，无需进行宏编程

13. 兼容性函数

编号	函　数	用　途
1	BETADIST	返回 beta 累积分布函数
2	BETAINV	返回指定 beta 分布的累积分布函数的反函数
3	BINOMDIST	返回一元二项式分布的概率
4	CHIDIST	返回χ^2分布的单尾概率
5	CHIINV	返回χ^2分布的单尾概率的反函数
6	CHITEST	返回独立性检验值
7	CONFIDENCE	返回总体平均值的置信区间
8	COVAR	返回协方差（成对偏差乘积的平均值）
9	CRITBINOM	返回使累积二项式分布小于或等于临界值的最小值
10	EXPONDIST	返回指数分布
11	FDIST	返回 F 概率分布
12	FINV	返回 F 概率分布的反函数
13	FTEST	返回 F 检验的结果

续表

编号	函　数	用　途
14	GAMMADIST	返回 γ 分布
15	GAMMAINV	返回 γ 累积分布函数的反函数
16	HYPGEOMDIST	返回超几何分布
17	LOGINV	返回对数累积分布函数的反函数
18	LOGNORMDIST	返回对数累积分布函数
19	MODE	返回在数据集内出现次数最多的值
20	NEGBINOMDIST	返回负二项式分布
21	NORMDIST	返回正态累积分布
22	NORMINV	返回正态累积分布的反函数
23	NORMSDIST	返回标准正态累积分布
24	NORMSINV	返回标准正态累积分布函数的反函数
25	PERCENTILE	返回区域中数值的第 k 个百分点的值
26	PERCENTRANK	返回数据集中值的百分比排位
27	POISSON	返回泊松分布
28	QUARTILE	返回一组数据的四分位点
29	RANK	返回一列数字的数字排位
30	STDEV	基于样本估算标准偏差
31	STDEVP	基于整个样本总体计算标准偏差
32	TDIST	返回学生 t-分布
33	TINV	返回学生 t-分布的反函数
34	TTEST	返回与学生 t-检验相关的概率
35	VAR	基于样本估算方差
36	VARP	计算基于样本总体的方差
37	WEIBULL	返回 Weibull 分布
38	ZTEST	返回 z 检验的单尾概率值

注：①“多维数据集函数“为 Excel 2007 中新加入的函数；

②所有的“兼容函数”已由新函数替换，这些新函数可以提供改进的精确度，其名称更好地反映其用法。虽然这些函数仍可向后兼容，但读者应该考虑从现在开始使用新函数，因为这些函数在 Excel 的将来版本中可能不再可用。有关新函数的详细信息，请查看“统计函数”部分；

③参见光盘文件“按类别列出的 Excel 函数.xlsx”或“按字母排序的 Excel 函数.xlsx”。

附录 B

概率基本知识

概率（probability）是某事件发生的可能性（likelihood）或机率（chance），前面章节中所指的假设检验获得的 P 值实际上就是概率值，事件 A 发生的概率通常记为 $P(\mathrm{A})$。概率的取值范围为 0～1，即 $P(\mathrm{A})\in[0,1]$；非 A 的概率为 $P(\bar{\mathrm{A}})=1-P(\mathrm{A})$。

联合概率（joint probability）（A 与 B 同时存在）为：

$$P(\text{A和B相交})=P(\mathrm{A}\cap\mathrm{B})=P(\mathrm{A}\mid\mathrm{B})P(\mathrm{B})$$

如果事件 A、B 相互独立（independent）（互不影响），则联合概率为：

$$P(\text{A和B相交})=P(\mathrm{A}\cap\mathrm{B})=P(\mathrm{A})P(\mathrm{B})$$

加法规则（addition rule）有：

$$P(\text{A或B的并集})=P(\mathrm{A}\cup\mathrm{B})=P(\mathrm{A})+P(\mathrm{B})-P(\mathrm{A})P(\mathrm{B})$$

当事件 A 与 B 为互斥（mutually exclusive）（有你无我）时，有：

$$P(\text{A和B相交})=P(\mathrm{A}\cap\mathrm{B})=0$$

$$P(\text{A或B的并集})=P(\mathrm{A}\cup\mathrm{B})=P(\mathrm{A})+P(\mathrm{B})$$

条件概率（conditional probability）是假定某事件出现的条件下，另一事件出现的概率。如已知事件 A、B 相互独立，在事件 B 出现的条件下，事件 A 出现的条件概率为：

$$P(\mathrm{A}|\mathrm{B})=\frac{P(\mathrm{A}\cap\mathrm{B})}{P(\mathrm{B})}=\frac{P(\mathrm{A})P(\mathrm{B})}{P(\mathrm{B})}=P(\mathrm{A})$$

一般地，条件概率可由贝叶斯定理（Bayes's theorem）求得，即贝叶斯公式为：

$$P(\mathrm{A}\mid\mathrm{B})=\frac{P(\mathrm{B}\mid\mathrm{A})P(\mathrm{A})}{P(\mathrm{B}\mid\mathrm{A})P(\mathrm{A})+P(\mathrm{B}\mid\bar{\mathrm{A}})P(\bar{\mathrm{A}})}$$

其中 $P(\mathrm{A})$ 为先验概率（prior probability）或边际概率（marginal probability）（没有考虑 B 的信息），$P(\mathrm{A}\mid\mathrm{B})$ 为后验概率（posterior probability）。

附录 C

Excel 2010 宏

如果经常在 Microsoft Excel 中重复某项任务，那么可以用宏自动执行该任务。宏是一系列命令和函数，存储于 Visual Basic 模块中，并且在需要执行该项任务时可随时运行。例如，如果经常在单元格中输入长度超过单元格默认宽度的文本字符串，则可以创建一个宏来将单元格格式设置为文本可自动换行。

下面通过一个示例来初步了解怎样利用宏解决单元格内文本自动换行的问题。B2 单元格内有一个很长的文本字符串（如图 C-1 所示），这看起来很不美观。

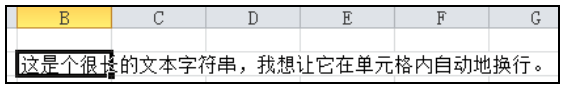


图 C-1　自动换行问题

而且如果右侧相邻的单元格 C2 内有内容（如图 C-2 所示），情况就变得更加糟糕，B2 单元格内的文本字符串不能完全显示出来。



图 C-2　相邻单元格有内容

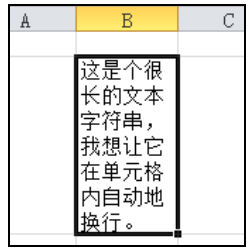
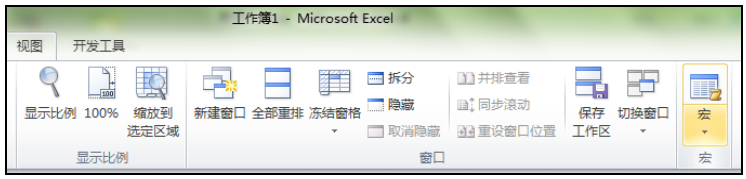


图 C-3　正确显示

为了解决这个问题，可以通过单击功能区内“开始→对齐方式→自动换行”命令，将单元格的格式设定为自动换行。于是，出现如图 C-3 所示的显示格式，这正是我们所需要的效果。但是，如果有大量的单元格内包含着这样的长文本字符串，而且这些单元格的分布又没有规律，按上述方法来处理就比较费时。这时，可以选择更为简便的方法，这就是宏。以下介绍各种有关宏的方法和具体的操作步骤。

1. 创建宏

单击“视图→宏”或“开发工具→宏”命令（如图 C-4 所示），出现“宏”对话框。



(1)

图 C-4　调用宏



(2)

图 C-4　调用宏（续）

在“宏名”栏内输入“AutoWrap”，如图 C-5 所示。

单击“创建”按钮，进入 VBA 编辑器，如图 C-6 所示。

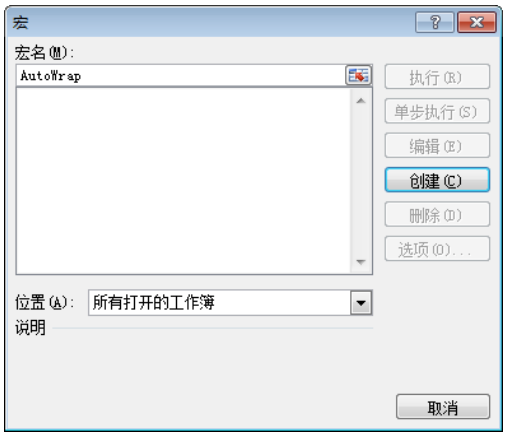


图 C-5　设定宏名

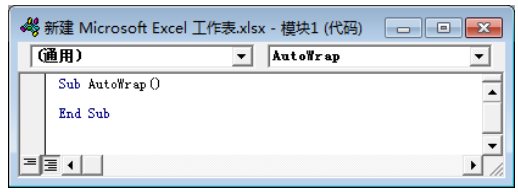


图 C-6　VBA 编辑器

VBA 编辑器共有 3 行，第 2 行是空行，将“Selection.WrapText = True”填入第 2 行的空白处。填完以后，宏的全部代码如下。

```
Sub AutoWrap()
Selection.WrapText = True
End Sub
```

编写完成的宏代码如图 C-7 所示。

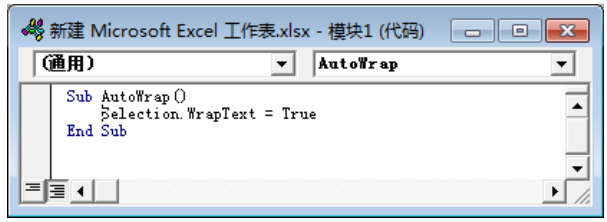


图 C-7　完成的宏

这一段宏代码的意思是说，将当前选中的单元格的格式设置为文本自动换行。至此，创建宏的基本过程就可以算结束了。但为了使这个宏今后用起来更方便，还需要完成一些辅助过程。

继续单击“视图→宏”命令，出现“宏”对话框，如图 C-8 所示。

单击“选项”按钮，出现“宏选项”对话框，如图 C-9 所示。

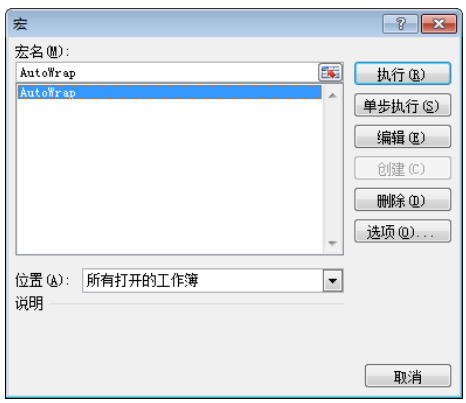


图 C-8　继续调用宏

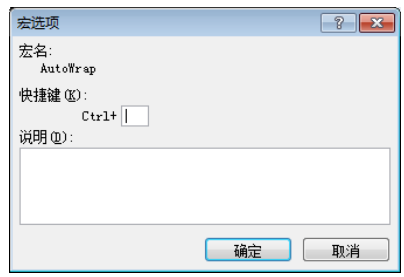


图 C-9　“宏选项”对话框

在“快捷键”栏中输入大写英文字母 A，也可以输入任何其他大写或小写英文字母；在“说明”栏中填入“当单元格内文本字符串的长度超过单元格的宽度时，文本将自动换行。”，当然，你也可以填入任何其他的你认为适当的文字。填完以后的“宏选项”对话框如图 C-10 所示。

单击“确定”按钮，又回到了刚才见过的“宏”对话框，注意，这时候你所看见的“宏”对话框的“说明”栏中出现了刚才输入的文字“当单元格内文本……”（如图 C-11 所示）。

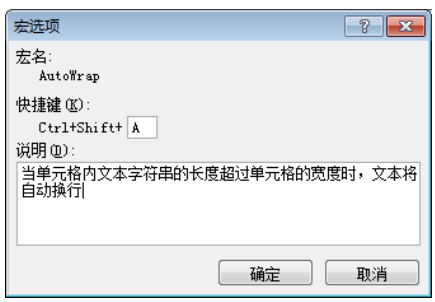


图 C-10　宏选项说明

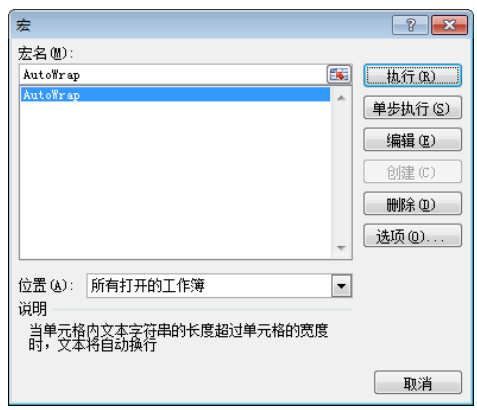


图 C-11　单击“确定”按钮后的宏说明

2. 调用宏

调用宏可采用如下方法。

（1）利用“宏”对话框调用宏

选择需要自动换行的文本所在单元格，如单元格 B2，单击功能区内“视图→宏”命令，在随后出现的“宏”对话框中选择宏名为“AutoWrap”的宏，单击“执行”按钮。这个方法比较笨拙。

（2）利用快捷键调用宏

选择需要自动换行的文本所在单元格，如单元格 B2，按 Ctrl+Shift+A 组合键即可。

方法（2）简单，但如果时间长了以后，调用宏的 Ctrl+Shift+A 组合键也许会忘记，此时可采用如下方法。

（3）利用图形对象调用宏

首先在工作表中创建一个图形对象，然后将宏指定给这个图形对象。需要调用宏时，只需用鼠标单击一下这个图形对象即可。

图形对象包括用绘图工具产生的任何图像，还包括“开发工具”中的“表单控件”、Active X 控件等。以下利用绘图工具创建图像。

①创建图形对象

单击“插入→形状”命令，如图 C-12 所示。

在“形状”命令弹出的菜单中，选择“基本形状→棱台”，如图 C-13 所示。

从 D4 单元格处开始向右下方拖曳鼠标，画出一个棱台，紧接着敲击键盘，输入文字“自动换行”，如图 C-14 所示。图形对象的创建即完成。

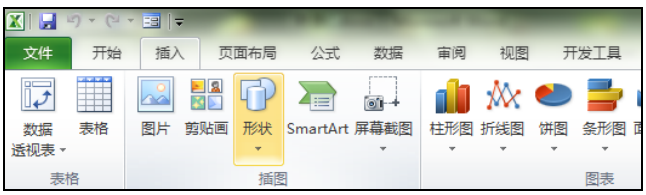


图 C-12　调用形状工具

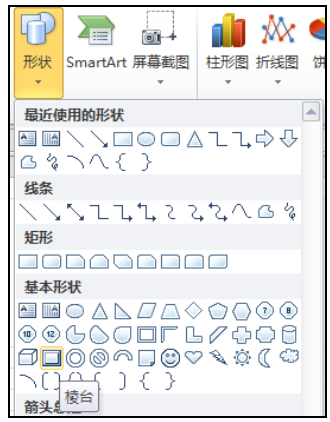


图 C-13　自选图形

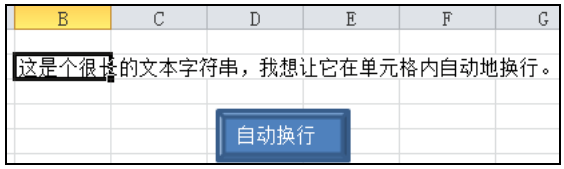


图 C-14　输入文字“自动换行”

②将宏指定给图形对象

刚才产生的图形对象暂时还没有任何功能。现在我们把前面已经创建的宏“AutoWrap”指派给这个图形对象，或者说，为这个图形对象指定一个宏“AutoWrap”。

用鼠标右键单击这个图形对象，在随后出现的快捷菜单中选择“指定宏”选项，出现“指定宏”对话框，如图 C-15 所示。

在“宏名”栏输入“AutoWrap”，在“宏名”栏内显示出你输入的文字“AutoWrap”，的同时，最底下的“说明”栏中也出现了一段文字，仔细一看会发现，这正是你在前面创建宏的过程中所给出的关于这个宏的功能的说明文本，如图 C-16 所示。

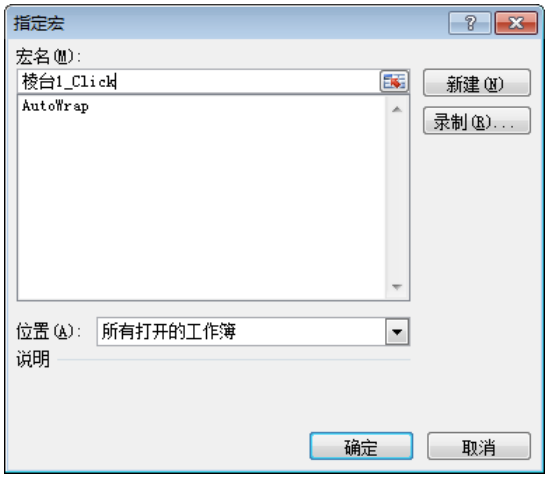


图 C-15　“指定宏”对话框

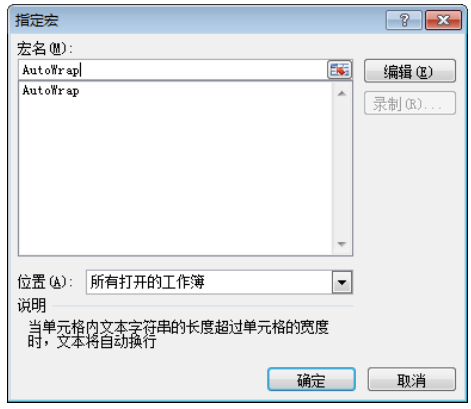


图 C-16　指定宏名

至此，方法（3）的操作完成，现在您可以利用图形对象来调用宏了。

如果有很多文件都需要调用这个 AutoWrap 宏，则可采用下面方法实现。

3. 录制宏

通过录制宏，将关于该过程的每一个操作步骤的信息变成一系列 Excel 可执行命令。然后，即可运行宏来重复所录制的过程或“回放”这些命令。如果在录制宏时出错，所做的修改也会被录制下来。Visual Basic 在附属于某工作薄的新模块中存储每一个宏。

4. 使宏易于运行

可以在“宏”对话框的列表中选择所需的宏，并运行宏。如果希望通过单击特定按钮或按下特定组合键来运行宏，可将宏指定给某个工具栏按钮、键盘快捷键或工作表中的图形对象。Excel 2003 是要先添加按钮再指定按钮的宏，而 Excel 2010 是在加载宏后将其视为内部的命令，然后直接添加进工具栏即可。

将宏添加到“快速访问工具栏”内按钮的操作为：在“文件→选项→快速访问工具栏”内“从下列位置选择命令”下拉列表中选择“宏”，然后在其下方的列表中选择“添加”宏“AutoWrap”（如图 C-17 所示），最后单击“确定”按钮。

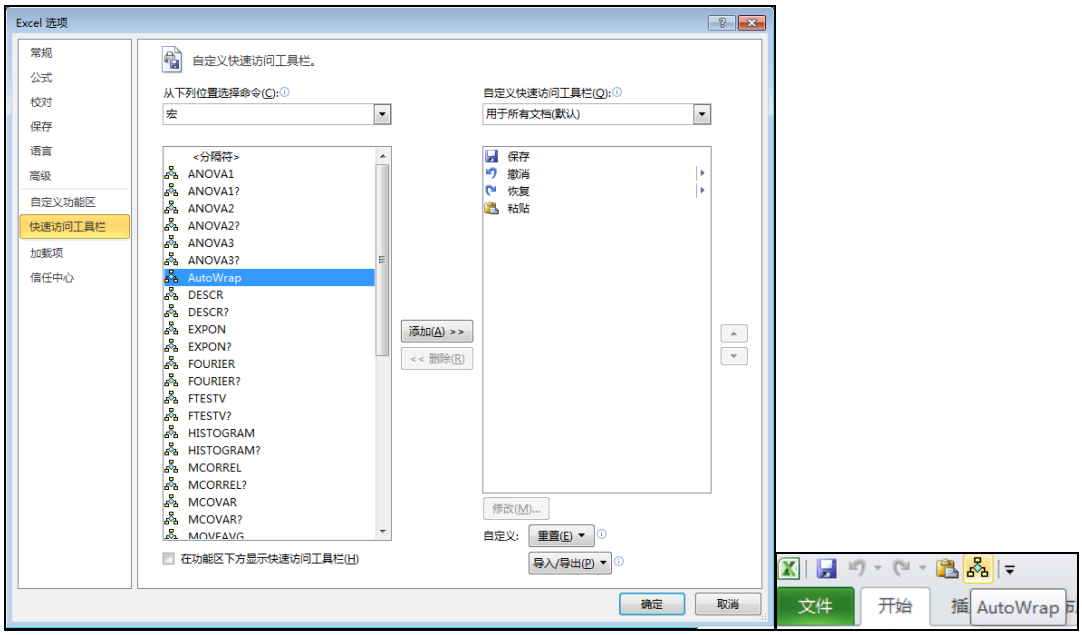


图 C-17　将宏添加到“快速访问工具栏”内按钮

将宏指定给某个键盘快捷键的操作为：“宏→选项”命令，参见图 C-8 至图 C-10。

5. 管理宏

宏录制完后，可以用 Visual Basic 编辑器查看宏代码，以进行改错或更改宏的功能。例如，如果希望将“文本换行的宏”还可以实现文本变粗体的功能，则可以再录制另一个将单元格文本变为粗体的宏，然后将其中的指令复制到用于文本换行的宏中即可。

“Visual Basic 编辑器”是一个为初学者设计的编写和编辑宏代码的程序，它提供了很多联机帮助。不必学习如何编程或如何用 Visual Basic 语言来对宏进行简单的修改。利用“Visual Basic 编辑器”，你可以编辑宏、在模块间复制宏、在不同工作簿之间复制宏、重命

名存储宏的模块或重命名宏。

6. 宏的安全性

Excel 对可通过宏传播的病毒提供了安全保护。如果你与其他人共享宏，则可使用数字签名来验证其他用户，这样就可保证其他用户为可靠来源。无论何时，打开某一个用户的包含宏的工作簿时，都可以先验证这个宏的来源，只有该用户被确认为可靠来源，Excel 才会启用宏。

为了能利用一般的 Excel 宏命令，应该选择“文件→选项→信任中心→信任中心设置→宏设置→禁用无数字签署的所有宏”，并单击“确定”按钮。

7. 应用实例

实例 1：见光盘文件“附录 1 统计图表的制作步骤.xlsm”，该宏文件演示了制作普通图形的方法和步骤。

实例 2：见光盘文件“附录 2 置信区间图制作.xlsm”，该宏文件演示了制作置信区间图形的方法和步骤。

实例 3：见光盘文件“附录 3 正态分布动态演示.xlsm”，该宏文件演示了正态分布图形的动态特点。正态分布有均数、标准差两个参数，该程序将这两个参数做成了两个按钮。在保持均数（或者标准差）不变的情况下，观察另一个参数对图形形状的改变情况（此程序在速度快的计算机上，图形可能不变动，大家可以观察左侧数据的变化）。

实例 4：见光盘文件“附录 4 二项分布动态演示.xlsm”，该宏文件演示了二项分布图形的动态特点。二项分布有试验次数和总体率两个参数，该程序将这两个参数做成两个按钮。在保持试验次数（或者总体率）不变的情况下，观察另一个参数对图形形态的改变情况（此程序在速度快的计算机上，图形可能不变动，大家可以观察左侧数据的变化）。

附录 D

Excel 2010 的数据分析新功能

附录 D Excel 2010 的数据分析新功能

Excel 2010 中包含了许多新功能，例如包括：自定义功能区外观；“文件”选项卡可方便访问文件与打印文件；Excel 中插入公式类似于 Microsoft Word 公式编辑器进行公式编辑器等。

在数据分析方面有下列新功能：

（1）在一个单元格内绘制迷你图；

（2）用于数据透视表分段和筛选的切片器；

（3）可根据多达 100 万行的数据（来源于数据库、电子表格、网站）做快速数据透视；

（4）改进后的规划求解可使之前老版本返回不正确答案情况下找到“最好”的解决方案；

（5）Excel 统计函数的精确度更高，并加入了几个新的功能（如 RANK.EQ、AGGREGATE、WORKDAY.INTL 和 NETWORKDAYS.INTL 等）。

（6）“数据条”图形更为优化。

下面简要举例说明 Excel 2010 的几个新特点：

1. 迷你图

迷你图是适合单元格的微型图表，图形有折线图、柱状图、盈亏图 3 种（如图 D-1 所示）。迷你图是放置在一个单元格的数据对应图形摘要。如打开光盘文件“15-4 重复测量曲线.xlsx”的“迷你图”工作表，绘制的 4 个时点某药新旧剂型血药浓度的迷你图如图 D-2 所示。



图 D-1 迷你图界面

	A	B	C	D	E	F	G	H
1	表 4个时点的某药新旧剂型血药浓度（μmol/L ）							
2			时间（小时）					
3	组别	编号	0	4	8	12	折线图	柱状图
4	旧剂型	1	90.53	142.12	65.54	73.28		
5		2	88.43	163.17	48.95	71.77		
6		3	100.01	144.75	86.06	80.01		
7		4	46.32	126.33	48.95	39.54		
8		5	73.69	138.96	70.02	60.89		
9		6	105.27	126.33	75.01	83.66		
10		7	86.32	121.06	78.95	70.24		
11	新剂型	8	70.53	97.38	112.12	58.5		
12		9	68.43	95.27	133.17	56.9		
13		10	57.37	78.43	83.16	48.34		
14		11	105.8	120.54	136.33	84.03		
15		12	80.01	104.75	114.75	65.61		
16		13	56.32	75.27	96.33	47.52		
17		14	53.69	110.02	138.96	45.44		
18		15	85.27	110.01	126.33	69.47		
19		16	66.32	115.27	129.06	55.29		

图 D-2 迷你图实例

2. 切片器

使用切片器可以实现数据透视表或筛选的分类方式显示。如打开光盘文件“03-04 妊娠分娩.xlsx”，先采用数据透视表以“职业”为行标签，“分娩方式”为列标签建立二维列联表（参见第 5 章），然后单击“插入→切片器”命令，选择“文化程度”，在选择其中的大学选项，所得的切片图如图 D-3 所示。这实际是 1402 个妊娠分娩孕妇中“大学”文化程度有 208 人，其中工人 15 人（剖宫产 6 人，顺产 9 人）。

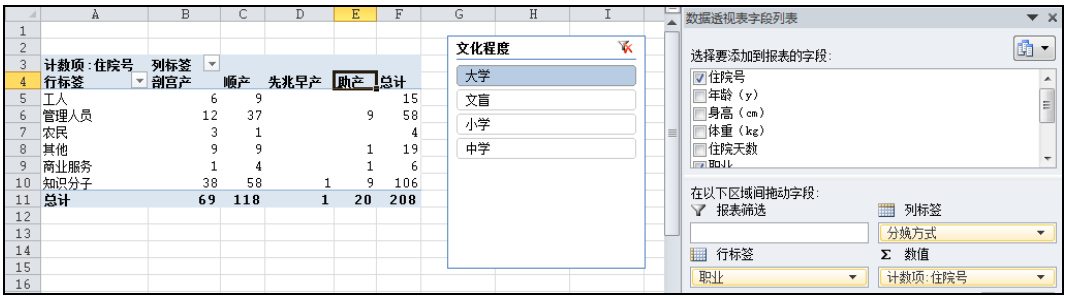


图 D-3　切片图实例

3. 公式输入

Excel 2010 公式编辑栏重新调整了尺寸，可以容纳长的公式而不会与工作表交迭并覆盖标题或数据。在公式栏中可以包含大约 8000 个字符和多达 64 层嵌套，而 Excel 2003 中只能包含 1000 个字符和 7 层嵌套。函数参数的数量也从 Excel 2003 的 30 个增加到了 255 个。功能区中的“公式”选项卡，提供了对需要创建和审核公式命令的快速参考，同时包括了容易使用的函数库，显示了主要的函数类别，让用户迅速选择所需要的函数。在“计算”组中，“计算选项”按钮允许用户控制公式的计算。公式自动完成一项新功能，能够帮助用户记住函数确切的名称和公式语法。在开始输入公式时，Excel 将显示相关的函数列表（以及命名区域和表引用），用户可以使用 Tab 键或鼠标选择所需公式；然后，Excel 帮助用户输入正确的参数，并建议定义的名称或结构引用。

4. 内置函数

自 2007 版本以后，Excel 扩展了函数库，包括了 7 个新的多维数据集函数，允许从 SQL 服务器分析服务中提取信息；分析工具包将函数整合到了主函数库中。同时，也添加了下面的通用函数。所有函数参见光盘文件“按类别列出的 Excel 函数.xlsx”或“按字母排序的 Excel 函数.xlsx”，下面举例说明几个函数。

- IFERROR（逻辑）：判别公式错误，而不是使用 IF 和 ISERROR。
- AVERAGEIF（统计）：有条件地求一个区域的平均值，与 SUMIF 和 COUNTIF 相似。
- SUMIFS（数学和三角）：与 SUMIF 相似，但可以使用多个判断条件。

- COUNTIFS（统计）：与 COUNIF 相似，但可以使用多个判断条件。
- AVERAGEIFS（统计）：与 AVERAGEIF 相似，但可以使用多个判断条件。
- RANK.EX（统计）：返回一个数字在数字列表中的排位。其大小与列表中的其他值相关。如果多个值具有相同的排位，则返回该组数值的最高排位。
- RANK.AVG 函数（统计）：返回一个数字在数字列表中的排位：数字的排位是其大小与列表中其他值的比值；如果多个值具有相同的排位，则将返回平均排位。
- AGGREGATE 函数（数学和三角）：返回列表或数据库中的合计。
- WORKDAY.INTL 函数（日期和时间）：返回日期在指定的工作日天数之前或之后的序列号。
- NETWORKDAYS.INTL 函数（日期和时间）：返回两个日期之间的所有工作日数，使用参数指示哪些天是周末，以及有多少天是周末。

5. 排序和筛选

Excel 2010 中的排序和筛选已增强了一些有用的功能。现在，可以通过多达 64 层级来排序数据，也能够通过单元格颜色、字体颜色、单元格图标、单元格值等来排序。事实上，排序命令将随着排序的数据而变化：如果单元格包含的是文本，则可以选择升序和降序；如果是日期，则可以是从最老的日期到最新的日期或相反；如果是数字，则可以是从大到小或从小到大。

可以执行更复杂和动态的筛选。例如，对于数字数据，可用的筛选包括等于、大于、大于平均值、前 10 项等；如果所选列包含文本，则可以选择开头是、结尾是和包含等筛选；如果筛选数据包含日期，则可以选择明天、本周、上月、下季度、今年等。自动筛选列表可以达到 10000 项，也可以通过值、单元格颜色、字体颜色或单元格图标来筛选，并且可以在表中重复应用排序和筛选条件。

6. 统计表与图

用户更容易创建、格式化和扩展 Excel 表格，具体体现在以下几个方面：

① 可以打开和关闭标题；

② 自动扩展所计算的列来包括所添加的行；

③ 在每个标题单元格中都包含一个自动的筛选按钮，可以快速地排序和筛选；

④ 结构引用允许用户在公式中使用列标题名称和指定的项目代替单元格引用；

⑤ 可以使用自定义公式和文本输入来汇总行；

⑥ 随着添加或删除行，表格样式将自动调整更新。

与表和其他对象一样，Excel 2010 统计图格式也基于文档主题，新图表的外观也包含特殊的效果，例如柔化边缘、阴影、倾斜和 3D 效果。另一个有用的功能是只需一次单击就能切换行列数据，并且可以保存图表格式和布局为模板，以应用到将来的图表中。另一

个主要的变化是 Excel 的图表引擎现在也在 Word 2010 和 PowerPoint 2010 中使用。Excel 工作表可更方便用作 Word 和 PowerPoint 的图表数据表，因此这些应用程序可以利用 Excel 所提供的功能，包括公式、筛选、排序和链接图表到外部数据源。

7. 条件格式

Excel 2010 中的条件格式引入了一些新颖的功能，如色阶、图标集和数据条，可以使用户以一种更可理解的方式使分析数据可视化。根据数值区域里单元格的位置，用户可以分配不同的颜色、特定的图标或不同长度阴影的数据条。

同时 Excel 2010 也提供了不同类型的通用规则，使之更容易创建条件格式。这些规则为“突出显示单元格规则”和“项目选取规则”。使用“突出显示单元格规则”，可以从规则区域选择高亮显示的指定数据，包括识别大于、小于或等于设置值的数值，或者指明发生在给定区域的日期。“项目选取规则”允许用户识别项目中最大或最小的百分数或数字所指定的项，或者指定大于或小于平均值的单元格。

此外，Excel 2010 中的条件格式取消了可以应用于单元格区域中条件格式数的限制。不像 Excel 2003 中只能识别三个条件，Excel 2010 中指定的条件格式数只受计算机内存的限制。为了帮助追踪这些大量的条件格式规则，Excel 2010 提供了“条件格式规则管理器”，可以创建、编辑、删除规则，以及控制规则的优先级。

另外，Excel 2010 还可以自己定制条件格式，定义自己的规则或格式。

附录 E

Excel 2010 VBA 应用

可以认为 VBA（Visual Basic For Application）是非常流行的应用程序开发语言 Vasual Basic 的子集。实际上，VBA 是“寄生于”VB 应用程序的版本，在提高用户的工作效率方面起到了很大的作用。

本节以配套光盘目录的制作为例，来展示一下 VBA 的优秀之处。

本书的配套光盘中共有 200 多个 Excel 文件，将其做成目录并添加链接，将是一个比较大而且烦琐重复的工作。但是，应用了 VBA 代码之后，这个工作就显得非常迅速了。

生成目录步骤如下：

① 先借用 WinRAR 压缩软件生成 Excel 文件列表的一个文本文件。即打开 WinRAR 文件，在 WinRAR 软件中找到 Excel 光盘文件所在的子目录，并选取需要生成目录的文件（按 Ctrl+A 键）。单击 WinRAR 软件的“工具→生成报告”命令，输入格式改为“文本”，报告包含只选“普通文件”，单击“确定”按钮即可生成获得“目录生成试验”下的“配套光盘文件.txt”文件。

② 建立“目录生成试验”下的“章节目录.txt”文件。

③ 打开“目录生成试验”下的“TryList.xlsm”文件，按 Alt+F11 组合键，呈现 Microsoft Visual Basic 界面，单击“运行→运行子过程/窗用户体”命令（或直接按 F5 键），在“TryList.xlsm”文件的主界面便生成了带有链接的所有 Excel 文件目录。

“模块 1”VB 文件的具体代码及解释如下：

```
Sub BuildFileLink()                '生成 B 列文件名和超链接
Dim hang As Long
Dim sfileLink As String
Dim iAs Long
Dim t As String
Sheets("sheet1").Select                                  '工作区选择
Range("a1").Value = "章节"
Range("b1").Value = "文件名"
Open ThisWorkbook.Path& "\配套光盘文件.txt" For Input As #1
                                                         '按照文本方式打开文件列表

For hang = 2 To 1 + 210              '记录的行号的范围（注意根据情况修改!）
Range("B" & hang).Select                                 '单元格选择
Line Input #1, t                                         '读取一行，赋值到 t
If t <> "" Then
Selection.Value = t                                      '单元格赋值
sfileLink = "配套光盘文件\" & t                           '此处修改文件最终的放置位置，
                                                         即目录文件所在目录的下级目录
                                                         “配套光盘文件”
ActiveSheet.Hyperlinks.Add Anchor:=Selection, Address:=sfileLink,
TextToDisplay:=t                                         '添加超链接
     End If
If EOF(1) Then Exit For                                  '到达文件尾部
Next
Close #1
```

```
End Sub

Sub zhangBuild()           '章节名称匹配
Dim hang As Long, i As Long
Dim zNo As String, lNo As String, lastNo As String
Dim nameZ As String, t As String
Dim checkState As Long                              '状态标志：0，未找到；1，找到；
                                                  2，超过
Sheets("sheet1").Select
hang = 2
Open ThisWorkbook.Path& "\章节目录.txt" For Input As #2

For i = 1 To 20
If EOF(2) Then Exit For
Line Input #2, nameZ
loca = InStr(1, nameZ, "章", vbTextCompare)
zNo = Mid(nameZ, 2, loca - 2)
checkState = 0
Do
Range("B" & hang).Select
If (Left(Selection.Value, 1) = "0") Then
lNo = Mid(Selection.Value, 2, 1)
     Else
lNo = Mid(Selection.Value, 1, 2)
     End If
If lNo = zNo Then
Range("A" & hang).Select
Selection.Value = nameZ
   With Selection.Interior      '单元格背景颜色操作
         .Color = 16777164
         End With
     If checkState = 1 Then
    With Range("A" & hang).Font                     '对单元格进行字体颜色操作
               .Color = -52
               End With
          End If
checkState = 1
     Else
checkState = checkState * 2
     If checkState = 2 Then Exit Do
     End If
hang = hang + 1
If hang > 10000 Then Exit Do                        '设置溢出，避免死循环
Loop
Next
Close #2
End Sub
```

为了方便读者练习，上述代码在配套光盘根目录下的“目录生成试验”目录中的TryList.xlsm文件中已包含。注意：对于不同的目录数量需要对行数（如210）做适当的修改！

参 考 文 献

[1] Hans P, Albert A, Born JD, and Chapelle JP. Derivation of a bioclinical index in severe trauma. Intensive Care Medicine 1985,11:186-191

[2] Dupont WD. Power and sample size calculations. Controlled Clinical Trials 1990,11:116-128

[3] Dupont WD. Power calculations for matched case-control studies. Biometrics 1988,44:1157-1168

[4] Gerald Keller. Applied Statistics With Microsoft Excel. Duxbury Press, 2000,12.（北京：机械工业出版社，2004）

[5] http://en.wikipedia.org/wiki

[6] Lemeshow S, Hosmer DW, and Klar J. Sample size requirements for studies estimating odds ratios or relative risks. Statistics in Medicine,1988,7(7):759-764

[7] Mick O'Neill & Peter Thomson. Australian Journal of Experimental Agriculture, 1998, 38 617-22

[8] Zhou XH, Obuchowski NA, and McClish DA. Statistical Methods in Diagnostic Medicine. John Wiley & Sons, Inc., New York, 2002

[9] 陈卉. Bland-Altman 分析在临床测量方法一致性评价中的应用. 中国卫生统计 2007, 24(3):308-315

[10] 方积乾主编. 卫生统计学（第六版）. 北京：人民卫生出版社. 2008

[11] 方亚. 样本含量的估计方法（第六章），见董时富主编，生物统计学. 北京：科学出版社，2002

[12] 郭绣花，罗艳霞. 定性资料的比较（第 8 章）. 见方积乾主编，生物医学研究的统计方法. 北京：高等教育出版社，2007

[13] 郭祖超. 第六章 圆形分布. 见郭祖超主编. 医用数理统计方法（第三版）. 北京：人民卫生出版社，1988. 122-140

[14] 胡冬梅, 姜潮, 刘启贵等. 应用集中度和圆形分布分析自杀的高发时间和季节. 中国卫生统计, 2004, 21(6):360-361

[15] 胡良平，周诗国，宇传华，王家春. 样本含量估计（第 15 章），见方积乾主编，生物医学研究的统计方法. 北京：高等教育出版社，2007

[16] 孔超, 刘元凤, 沈续雷. 灰色预测模型的 SAS 程序改进. 中国卫生统计. 2008, 25(6)：

640-641
[17] 施学忠, 杨永红, 赵耐清. 两样本定量资料的比较（第 6 章）. 见方积乾主编，生物医学研究的统计方法. 北京：高等教育出版社，2007
[18] 孙振球主编. 医学统计学. 北京：人民卫生出版社 2002
[19] 王玖，徐天和，高永，石德文. 多组定量资料（第 7 章）. 见方积乾主编，生物医学研究的统计方法. 北京：高等教育出版社，2007
[20] 严明, 廖铁军. 基于 Excel 的灰色预测模型在土地利用规划中的应用. 安徽农业科学, 2007, 35(12) : 3627-3628
[21] 张福林，杨树勤．平均角的可信区间估计．中国卫生统计，1992，9(2)：20
[22] 张照寰, 曹素华. 第二十四章 圆形分布资料分析. 见金丕焕, 陈峰主编. 医用统计方法（第三版）. 复旦大学出版社，2009. 327-334

/ 相关图书 /



《Excel 2010操作与技巧》
作者：王建发 李术彬 黄朝阳
ISBN：978-7-121-12043-5
出版日期：2010年11月



《Excel 2010函数与公式》
作者：陈国良 荣胜军 黄朝阳
ISBN：978-7-121-12042-8
出版日期：2010年11月



《Excel2010 数据透视表大全》
作者：黄朝阳 荣胜军 陈树青
林盘生 薛宇锋
ISBN：978-7-121-18959-3
出版日期：2013年1月

Broadview®
www.broadview.com.cn
博文视点·IT出版旗舰品牌
技术凝聚实力·专业创新出版

Excel与数据分析（第3版）

内容概要

·数据整理篇

Excel基本知识
公式与函数
数据清单的建立与利用
数据的编辑与查询
分类汇总与透视分析

统计分析篇·

统计图表
统计描述
区间数据的假设检验
名义与有序数据的假设检验
相关与回归

·高级应用篇

变量求解与方案优选
预测分析
寿命表
随机化分组与样本含量估计
高级科研统计图与Meta分析
诊断试验评价

电脑实验篇·

概率分布
随机抽样与抽样误差
假设检验的概率分布与统计用表
其他电脑实验与置信区间估计

上架建议：数据分析 > Excel

ISBN 978-7-121-20739-6

定价：79.00元（含光盘1张）

新浪微博 weibo.com @博文视点Broadview

策划编辑：张春雨　符隆美
责任编辑：付　睿
封面设计：侯士卿